ANALYSIS OF SPEECH AND OTHER SOUNDS

A thesis presented for the degree of
Doctor of Philosophy
in Electrical & Electronic Engineering,
at the
University of Canterbury,
Christchurch, New Zealand.

by
Cornelius William Thorpe
B. E. (Hons 1)
November 1990

Abstract

This thesis comprises a study of various types of signal processing techniques, applied to the tasks of extracting information from speech, cough, and dolphin sounds.

Established approaches to analysing speech sounds for the purposes of low data rate speech encoding, and more generally to determine the characteristics of the speech signal, are reviewed. Two new speech processing techniques, shift-and-add and CLEAN (which have previously been applied in the field of astronomical image processing), are developed and described in detail. Shift-and-add is shown to produce a representation of the long-term "average" characteristics of the speech signal. Under certain simplifying assumptions, this can be equated to the average glottal excitation. The iterative deconvolution technique called CLEAN is employed to deconvolve the shift-and-add signal from the speech signal. Because the resulting "CLEAN" signal has relatively few non-zero samples, it can be directly encoded at a low data rate. The performance of a low data rate speech encoding scheme that takes advantage of this attribute of CLEAN is examined in detail. Comparison with the multi-pulse LPC approach to speech coding shows that the new method provides similar levels of performance at medium data rates of about 16kbit/s.

The changes that occur in the character of a person's cough sounds when that person is afflicted with asthma are outlined. The development and implementation of a micro-computer-based cough sound analysis system, designed to facilitate the ongoing study of these sounds, is described. The system performs spectrographic analysis on the cough sounds. A graphical user interface allows the sound waveforms and spectra to be displayed and examined in detail. Preliminary results are presented, which indicate that the spectral content of cough sounds are changed by asthma.

An automated digital approach to studying the characteristics of Hector's dolphin vocalisations is described. This scheme characterises the sounds by extracting descriptive parameters from their time and frequency domain envelopes. The set of parameters so obtained from a sample of click sequences collected from free-ranging dolphins is analysed by principal component analysis. Results are presented which indicate that Hector's dolphins produce only a small number of different vocal sounds. In addition to the statistical analysis, several of the clicks, which are assumed to be used for echo-location, are analysed in terms of their range-velocity ambiguity functions. The results suggest that Hector's dolphins can distinguish targets separated in range by about 2cm, but are unable to separate targets that differ only in their velocity.

Acknowledgements

Many people have contributed to the completion of this thesis by way of their support, encouragement, and advice. I am indebted to each of them.

In particular, I thank Professor Richard Bates for the great amount of energy he expended as my supervisor, both in the ideas he contributed during the course of my research, and in his efforts to impose literary perfection on this thesis.

I would also like to thank Bill Kennedy for his help with many practical aspects of my research. Richard Fright made many invaluable contributions, both while he was a post-doctoral fellow in this department, and at Christchurch Hospital where he was instrumental in getting the cough analysis project underway.

During the period of my study I have had the opportunity to work with many people, and I thank all those people for their help, interest, and stimulation. I would especially like to thank each of Andrew Elder, Andrew Rolls, Bruce Davey, Catherine Watson, Martin Clark, Nigel Brieseman, Peter Gardenier, Qing Wu, Richard Lane, and Tracy Clark for many interesting discussions and much helpful advice.

In addition to the engineering staff and fellow students mentioned above, I am grateful to Drs. Les Toop and Ken Dawson of the Christchurch School of Medicine, and Steve Dawson of the Zoology Department at the University of Canterbury. Without these people I would not have had the opportunity to broaden my research into the areas of cough and Hector's dolphin sounds. I thank them each for their willingness to collaborate in this research, for their help during the course of our joint research, and for their essential advice on the non-technical content of the relevant chapters in this thesis.

I acknowledge the receipt of a New Zealand University Grants Committee Post-graduate Scholarship, a Christchurch City Council Electricity Department Research Scholarship, and a Masonic Fellowship in Paediatrics.

Finally, I wish to thank my family, friends, and flatmates who have supported me and helped make my life bearable (if not enjoyable). Many have had to contend with my (pitiful) excuse of "Wait until my thesis is finished...". I am glad to say now that it finally is.

Contents

Preface

Much of modern engineering science is concerned with the processing and analysis of data obtained from quantitative observations of phenomena. The advent of the digital computer has led to an explosion in the complexity and quantity of the information processing that is now feasible. As computers become increasingly powerful, miniaturised and reduced in cost, it is realistic to think of applying them to problems that were previously considered too computationally complicated or insufficiently cost effective. This thesis details the development and implementation of algorithms for harnessing the computational power of the computer to the task of extracting useful types of information from various types of sounds, namely speech and coughs of humans, and dolphin vocalisations.

Professor R.H.T. Bates' research group here at the Department of Electrical and Electronic Engineering of the University of Canterbury has, for the last two decades, been active in applying computer-based information processing techniques to a wide range of engineering and scientific topics (Bates, 1987). Research has ranged from various inverse problems (notably those related to computed tomography and ultrasonic imaging), through image restoration studies both fundamental (phase retrieval and blind deconvolution) and applied (to photography, astronomy, microscopy and generalisations of holography), to applied electromagnetics, various biomedical problems, the application of computers to human-machine interaction and the extraction of information from speech.

The speech group began in the late 1970s, growing out of research into computer-controlled aids for musicians (Tucker *et al.*, 1977) and methods of extracting pitch information from speech and music sounds (Tucker and Bates, 1978; Brieseman, 1984). That research led to investigations into micro-computer based speech aids for disabled people, aiming especially for real-time operation with the use of digital signal processor (DSP) integrated circuits (Turner, 1986). Current research includes the development of micro-computer based tools for speech therapists, techniques of low data rate speech encoding and the development of algorithms for reliable word and speaker recognition (Bates *et al.*, 1987). As well as employing standard speech analysis techniques, this research has benefited from the novel (to speech processing) techniques introduced by Professor Bates from his wide experience in other areas of information processing.

I entered the world of information processing in 1986 and since then have been fortunate to have had the opportunity to work on three separate projects, each of which involves a different facet of information processing. The first, and major, part of my research has been in the area of speech analysis. I began by working on the implementation of a new speech analysis technique called shift-and-add (SAA). This was developed in Professor Bates' astronomical imaging group, and is a species of blind deconvolution (Bates, 1982). SAA, originally applied to speech signals by Nigel Brieseman (a PhD. student at the time), is a technique whereby an estimate of the average (glottal) excitation of a speaker can be extracted in a simple fashion from a speech signal (Brieseman *et al.*, 1987).

The aim of my research was to employ the SAA technique in such a way that the naturalness of speech produced by low data rate encoding schemes could be improved. I began by investigating several methods of removing the effect of the SAA signal (which is assumed to approximately represent the average glottal excitation, see Chapter 4) from the speech signal in order to improve the performance of techniques such as linear predictive coding (LPC). LPC modelling is predicated upon the assumption that the speech signal can be represented by a convolution between a random, gaussian, excitation signal and an all-pole filter (Itakura and Saito, 1968). Because this assumption is violated by (voiced especially) speech signals, difficulties are sometimes experienced in extracting reliable LPC parameters from speech signals. Parameter estimation can be improved by removing "zeros" from the speech spectrum, and this is what I initially aimed to accomplish.

At the suggestion of Professor Bates and the continued insistence of Dr. Richard Fright (then a postdoctoral fellow in the Department) I implemented the subtractive deconvolution technique called "CLEAN", another import from (radio-)astronomical imaging. At first the results did not seem encouraging, since the result of a CLEAN deconvolution is a signal containing many "spikes" (which can nevertheless be a consistent outcome of the deconvolution of one continuous signal from another). However, it soon became clear that the number of non-zero samples that were adequate to represent the signal (as far as listening to the reconstructed speech was concerned) was relatively small, and that speech signals could be encoded at low data rates merely by appropriately combining the SAA signal and the non-zero CLEAN pulses (Thorpe and Bates, 19XX). Much of the work that I have undertaken in the speech processing field involves the development of techniques to improve the quality of the reconstructed speech whilst restricting the number of non-zero pulses (and hence the data rate).

There are many different methods of encoding speech at low data rates, with each technique offering its own data rate/quality/complexity tradeoffs. SAA/CLEAN is a low-complexity coder that falls somewhat in between the high-quality, medium–high data rate waveform coders and the low data rate vocoders. It is akin to the multi-pulse LPC (MP-LPC) technique (Atal and Remde, 1982) in its use of a sparse pulse sequence to help represent the speech signal. However, whereas MP-LPC employs a LPC filter as an estimate of the short-term spectral content of the signal, SAA/CLEAN uses the SAA signal as an estimate of the long-term "average" component of the speech signal. Because of this, the CLEAN pulses must represent more of the short-term structure of the speech signal than is the case in MP-LPC.

The second research area that I became involved in arose in late 1987 when Dr. Les Toop, a senior Lecturer in Community Health at the Christchurch School of Medicine, approached Professor Bates for advice on analysing the sounds of coughs produced by children who might have asthma. Dr. Toop had spent some time at Edinburgh University researching the occurrence of cough and asthma, and was keen to investigate the differences between the various types of coughs. Since asthmatic and non-asthmatic coughs "sounded different", it seemed reasonable that some computer-based analysis techniques could be devised that would characterise the differences. After preliminary spectral analyses (performed on the EEE Departmental VAX computer) indicated that there were differences between asthmatic and non-asthmatic cough sounds (Toop *et al.*, 1989a), we decided to embark on a clinical trial to compare the analyses with standard asthma diagnostic tests. To this end, I spent my Thursdays for the next 18 months developing a micro-computer based cough sound collection/display/analysis system (Thorpe *et al.*, 19XXa). This work afforded me much experience in the art of computer programming, especially in regard to its real-time, multi-tasking, user-interfacing and data analysis and management aspects.

The third part of my research has been to do with the sounds emitted by Hector's dolphin, a small dolphin native to New Zealand. I was introduced to this topic by Stephen Dawson, a PhD. student in the Zoology Department of the University of Canterbury, who had spent several years studying the dolphins in their natural habitat (which is, needless to say, far more pleasant than that afforded to electrical engineering students!). As part of his study, Steve had recorded many hours of their vocalisations and was interested in applying computer-based analysis techniques to try and characterise the sounds. Together with Drs. Richard Fright, Kathy Garden and Peter Gough, I looked at various ways in which the "clicks" could be characterised and identified. It soon became clear that there was no obvious single descriptor that would characterise them.

Zoologists are very interested in examining the structure of sounds produced by animals, and comparing the repertoire of an animal with its behavioural patterns in order to understand them more thoroughly. Up until recently, the only techniques available to analyse the sounds were spectrographic analysis or subjective listening. Statistical techniques were often employed, with variables being estimated visually from the spectrograms. Several recent studies have, however, employed digital signal processing techniques to quantitatively extract features from the sounds (cf. Clark, 1982).

Together with Steve Dawson and Richard Fright, I developed a variety of procedures for extracting different types of features that we thought could be relevant to characterise the clicks. By comparing these features with the behavioural information that Steve had recorded with the sounds, we hoped to discover if any types of sounds were employed in particular circumstances (Dawson and Thorpe, 1990). By means of a high-level signal processing language on our VAX computer we were able to quantitatively extract descriptive features from over 400 examples of the sounds (Thorpe and Dawson, 19XX), a task that would have been virtually impossible if Steve had employed the traditional technique of estimating variables from the spectrograms by hand and eye. During the course of this project, I gained much experience in the techniques and pitfalls (!) of attempting to classify groups of data according to (automatically obtained) estimates of their characteristic features.

In an attempt to extract further information from the clicks, and after much encouragement from Professor Bates (who worked in the radar field up until a quarter of a century ago), I implemented a program to calculate the ambiguity surface of a dolphin's sonar pulse (Thorpe *et al.*, 19XXb). The ambiguity surface provides an indication of the intrinsic range and velocity resolving capabilities of the echo-location signal, and hence of the dolphin who is employing that signal.

The chapters of this thesis are summarised in the following paragraphs, which also identify my original research contributions.

Chapter 1 summarises signal processing concepts and techniques. The mathematical basis on which signals are represented is introduced. The Fourier and z-transforms are defined and some of their properties which pertain to signal processing are developed. Some techniques employed in signal processing, such as spectral estimation, deconvolution and sampling are discussed, followed by a brief introduction to the types of sounds encountered in the biological world.

In Chapter 2 an introduction to speech sounds and the models employed in their analysis is presented. The phonetic and linguistic approaches to speech analysis are described in order to shed some light on various problems encountered in the development of speech analysis, synthesis and recognition algorithms. The physiology of the speech production and perception mechanisms is discussed, as are several models of these processes.

Chapter 3 is a review of established methods of speech analysis. Techniques covered include pitch estimation, voiced/unvoiced decision, various forms of spectral analysis and linear prediction of speech. An overview of various methods of low data rate speech encoding, glottal pulse shape estimation, speech and speaker recognition, and other applications of speech analysis technology is also presented.

The new technique of shift-and-add (SAA) is described in Chapter 4. A brief introduction to the astronomical origins of this algorithm is followed by an in depth examination of its application to speech signals. Apart from the introduction, the basic speech SAA algorithm, and the iterative refinement scheme described in §4.3.3, all the material presented in this chapter relates to my original research.

Chapter 5 presents the original research that I have done in the area of low data rate speech encoding. I discuss the "CLEAN" method of subtractive deconvolution, both in its original astronomical setting and as I have applied it to speech signals. The features of CLEAN which lead to a method of low data rate speech encoding are explained, as are some of the practical aspects of encoding speech at low data rates while retaining adequate speech quality. Results are presented and the method is compared in detail with the multi-pulse LPC approach to speech encoding.

Chapter 6 contains a description of the micro-computer based cough sound analysis system that I developed. As with the speech sounds, a brief introduction is provided to the physiology of coughs and the mechanism of sound production. Details of the implementation of the analysis system are presented together with an exposition of the analysis techniques employed and preliminary results obtained thereby. Apart from the introductory sections, all the material presented in this chapter is the result of my original research.

Chapter 7 describes the analysis techniques employed to characterise and study the sounds recorded from Hector's dolphins. The chapter includes descriptions of both the feature analysis and measurement scheme, and the statistical techniques employed to analyse the structure of the resulting data set. An analysis of the echo-location capability of the sounds is also presented. All the material in this chapter, apart from the background section and the zoological aspects of the work (which includes some of the statistical analyses and their interpretations), relates to my original research.

Chapter 8 offers specific conclusions on each of the three areas studied as well as some more general conclusions about analysing and characterising signals. Suggestions for further research in each area are also made.

Publications and conference presentations prepared during the course of my research are listed below.

Articles published and in press.

BRIESEMAN, N.P., THORPE, C.W. and BATES, R.H.T. (1987), 'Nontactile estimation of glottal excitation characteristics of voiced speech', *IEE Proceedings Pt. A*, Vol. 134, No. 10, December, pp. 807–813.

TOOP, L.J., THORPE, C.W. and FRIGHT, W.R. (1989), 'Cough sound analysis: A new tool for the diagnosis of asthma?', *Family Practice*, Vol. 6, No. 2, May, pp. 126–128.

TOOP, L.J., DAWSON, K.P., and THORPE, C.W. (in press), 'A portable system for the spectral analysis of cough sounds in asthma', *J. Asthma*, Vol. 27, No. 2, December 1990 (to appear).

DAWSON, S.M. and THORPE, C.W. (in press), 'A quantitative, digital analysis of the acoustic repertoire of Hector's dolphin', *Ethology*, December 1990 (to appear).

THORPE, C.W. and DAWSON, S.M. (in press), 'Automatic measurement of descriptive features of Hector's dolphin sonar signals', *Journal of the Acoustical Society of America*, December 1990 (to appear).

Articles in preparation and revision.

THORPE, C.W., DAWSON, K.P., and TOOP, L.J. , 'Spectrographic analysis of the coughs of three childhood respiratory diseases', *Journal of Paediatrics and Child Health*, In revision.

THORPE, C.W., TOOP, L.J., DAWSON, K.P., and FRIGHT, W.R. , 'A micro-computer based interactive cough sound analysis system', *Computer Methods and Programs in Biomedicine*, Submitted for publication.

THORPE, C.W., BATES, R.H.T. and DAWSON, S.M. , 'Intrinsic echolocation capability of Hector's dolphin, *Cephalorhynchus hectori*', *Journal of the Acoustical Society of America*, In revision.

THORPE, C.W. and BATES, R.H.T. , 'Speech analysis/companding/resynthesis by shift-and-add and Clean', *IEEE Transactions in Acoustics, Speech and Signal Processing*, In revision.

Conference presentations.

DAWSON, S.M. and THORPE, C.W. (1989), 'Hector's dolphin sounds; a digital signal processing approach', *Eighth Biennial Conference on the Biology of Marine Mammals*, Pacific Grove.

TOOP, L.J., DAWSON, K.P. and THORPE, C.W. (1989b), 'Spectrographic analysis of cough sounds in asthma', *European Annals of Allergy and Clinical Immunology*, Supplement 5, p. 14.

WATSON, C.I., CLARK, T.M., ELDER, A.G. and THORPE, C.W. (1988), 'Multifarious real-time speech processing applications', In *Proc. NELCON*, Christchurch, September, pp. 65–70.

ELDER, A., BATES, R.H.T., BRIESEMAN, N.P., CLARK, T.M., FRIGHT, W.R., GARDEN, K.L., KENNEDY, W.K., SQUIRES, P.L., TURNER, S.G. and THORPE, C.W. (1987), 'Real-time speech therapy aid', In *Proc. NELCON*, Auckland, 1–3 September, pp 115–118.

BATES, R.H.T., BRIESEMAN, N.P., CLARK, T.M., ELDER, A.G., FRIGHT, W.R., GARDEN, K.L., KENNEDY, W.K., SQUIRES, P.L., THORPE, C.W., TURNER, S.G. and JELINEK, H.J. (1987), 'Interactive speech-defect diagnostic/therapeutic /prosthetic aid', In LETELLIER, J.P. (Ed.), *Real Time Signal Processing X*, Proceedings of SPIE - The International Society for Optical Engineering, 20-21 August, pp. 131–139.

DAVEY, B.L.K. and THORPE, C.W. (1987), 'Image and signal reconstruction by shift-and-add', In *IPENZ Conference Proceedings.*, Christchurch, May, pp. 147–157.

THORPE, C.W., BRIESEMAN, N.P., SQUIRES, P.L. and BATES, R.H.T. (1986), 'Estimating glottal excitation by digital processing of recorded speech', In *New Zealand Medical Journal*, Proceedings of the Christchurch Medical Research Society, 16 July 1986, November, pp. 910–911.

Chapter 1

Preliminaries

Many activities entail the observation and measurement of phenomena, followed by suitable processing of the measured data to extract useful information about the identity, behaviour and/or origins of the phenomena. For example, our senses can be considered as devices which measure patterns of light, sound, temperature or pressure, or the presence of certain chemicals in the air or in our food. Our brain then processes the measurements and thereby infers useful knowledge about the world around us. Scientific endeavour provides another example. It can be considered as the process of proposing a theory to account for some phenomenon of interest, followed by experiments in which observations and measurements of the phenomenon are made. The theory can then be refined or altered on the basis of analyses of the measured data.

In all activities of the kind implied above, accurate measurement of the desired phenomena is vital if useful information is to be obtained. Much effort in engineering science is devoted to the art of either measuring physical phenomena, or deducing their characteristics by indirect means because the phenomena themselves are physically inaccessible or otherwise not amenable to direct measurement. For example, attributes of the heavenly bodies can only be inferred by measuring, at a distance, the radiation that they emit. Analysis of this radiation can then provide (for example) estimates of the chemical composition of the heavenly bodies. Often a phenomenon is most conveniently measured by an indirect process. For example, changes in temperature are usually "measured" by recording the expansion or contraction effects that they have on a substance such as mercury.

When phenomena are observed indirectly or at a distance, a great deal of processing is often required, both to extract the required data from the measurements actually made and to compensate for adverse effects of particular conditions under which the observations are carried out (cf. Bates and McDonnell, 1986; Davey, 1989). For instance, the surfaces of a large satellite telecommunication antenna must be maintained at their specified shape to a high degree of accuracy if the performance of the antenna is not to suffer (Morris, 1985). Direct measurement of the surface shape is often inconvenient because of the large size of such an antenna, but by measuring the antenna's radiation pattern in an appropriate way (e.g. by making use of suitable instrumentation in a communications satellite), and by the application of suitable processing methods, the shape of the antenna's surface can be deduced (Gardenier *et al.*, 1986).

Whenever one observes a phenomenon quantitatively, the measurements are corrupted to some extent, either by unavoidable inaccuracies and distortions introduced by the measurement process, or by interference from unwanted phenomena. In the example of antenna measurement mentioned above, corruption occurs because of inadequacies in the measurement apparatus, distortion introduced by the passage of the radiation through the atmosphere, and possible unwanted contributions to the

measurements from other sources of electromagnetic radiation.

Sometimes the corruption is systematic and can be counteracted by operating on the data with the inverse of the corruption. Other types of corruption are stochastic and cannot be undone, although their effects can on occasion be minimised by, for instance, appropriate averaging techniques.

After suitable measurements of a phenomenon have been made, they must be analysed in order to extract the useful information inherent within them. The usefulness of any particular item of information embedded in a measurement depends upon the reason for which the measurement is performed. In addition, the amount of information conveyed by a particular measurement varies according to its novelty (cf. §1.1.5). For example, a cat may be interested in knowing about other animals near by and about certain other events occurring in its vicinity. By continually subconsciously analysing its "measurements" of the sounds that surround it while it naps, it can deduce the behaviour of sound-producing animals and events. However, only changes in the character of the sounds (such as those caused by sudden noises) provide new information for the cat. Each time a change occurs in the surrounding sounds, the cat must process the new information and ascertain its implications. In addition, the particular sounds that the cat is more interested in, such as the sound of the fridge door opening, or that of a bird scratching in the garden, are intermingled with the sounds from other (less interesting) sources.

The purpose of an analysis system is to extract desired types of information from measurements, discarding that which appears extraneous, and then to make some judgement of or response to the information thereby gained. The process of extracting information from a set of measurements entails the identification of patterns in the measured data. Practitioners of virtually all professions and sophisticated endeavours attempt to master the art of reliably identifying patterns hidden within sets of measurements. For instance, economists attempt to identify patterns in economic data that may help them develop and refine models of the economy. Another example is that of a grader in an orchard, who classifies fruit according to particular patterns in the shape and colouring of the fruit.

The cat, the economist, and the orchardist also illustrate the two types of analysis techniques which may be employed to extract information from measurements. Firstly, a theory or model of the phenomenon may be constructed and refined by analysis of observations of the phenomenon. A model allows the underlying mechanisms of the phenomenon to be understood. Furthermore, the model describes the relationship between the measured data and the desired information. Therefore, (and secondly) the model can be employed as a basis on which analysis of the measurements can be performed. By structuring the model in terms of certain specific types of information (e.g. information about the behaviour or identity of the phenomenon), the analysis extracts only those particular kinds of information from the measured data. This type of analysis can be used to remove redundant information (i.e. information that is irrelevant to the model) from the measured data. The analysis can also be structured so that a decision about each set of measurements can be made, based on a particular type of information embedded in the measured data. For example, the fruit grader mentioned above has a model of what acceptable and unacceptable fruit looks like. The acceptability of a particular fruit is inferred by analysing observations of the fruit in terms of the model.

The analysis techniques described in later chapters of this thesis are concerned with extracting information from measurements of certain sounds. In each case, a different type of information is required and so different analysis techniques are employed. Chapters 2 through 5 are concerned with techniques which attempt to reduce the re-

dundancy in speech sounds by extracting only the information essential to the speech message itself. To this end, the sounds are analysed according to models of how humans produce and perceive speech. Chapter 6 introduces a method of analysing cough sounds which is designed to ascertain the presence or otherwise of asthma in the person coughing. The sounds are analysed in such a way that reveals patterns in the sound that may be caused by the asthmatic condition. Finally, Chapter 7 describes the analysis techniques employed to determine the repertoire of sounds used by Hector's dolphins, which are native to New Zealand. The efficacy of the sounds for echo-location is also assessed. In the analysis of vocal repertoire, the information in the sounds is classified according to a set of characteristic features. The analysis aims to identify the features which are most important for characterising the differences between different classes of sounds. The analysis of echo-location efficacy proceeds by, first, assuming a certain model of how the dolphins process the sounds and, then, describing the information in the sounds in terms of that model.

The remainder of this chapter lays the groundwork for the above-mentioned analysis techniques. Some basic concepts in the measurement and analysis of phenomena are presented in §1.1. §1.2 summarises the mathematical concepts and techniques employed in the analyses, while §1.3 outlines some of the more practical aspects of applying the mathematical techniques to the analysis of measured data. Finally, §1.4 briefly introduces some of the characteristics of sounds, describes those which are biological in origin, and introduces the speech sounds that are used in the speech analysis techniques described in later chapters of this thesis.

1.1 Introduction to signals and systems

In this section, some of the concepts underlying signal analysis techniques are introduced. The discussion in §1.1.2 and §1.1.3 introduces the concepts of *signals* and *systems*, respectively, and describes how they are used to mathematically represent phenomena and measurements of phenomena. §1.1.4 introduces the concept of *noise*, which is important because it is present in all measurement situations (to a greater or lesser extent). Finally, the concept of *information* is defined in §1.1.5. Detailed explanation of the topics treated in this section can be found in many Electrical Engineering textbooks, such as those by Stremler (1982) Oppenheim and Willsky (1983), Haykin (1983), and Rorabaugh (1986).

1.1.1 Mathematical nomenclature

Before discussing mathematical modelling of phenomena, it is appropriate to introduce certain notation and terminology. Complex numbers have long been employed as a tool by mathematicians and scientists to represent quantities (such as $\sqrt{-1}$) that cannot exist in reality (Kreyszig, 1979). A complex number can be thought of as representing a point on the *complex plane*, which has *real* and *imaginary* axes. If the real and imaginary coordinates of a complex number c are a and ib, respectively, where $i = \sqrt{-1}$, then c is written as

$$c = a + ib. \tag{1.1}$$

The real and imaginary parts of a complex quantity are also denoted by the notation:

$$\begin{aligned} a &= \mathcal{R}\{c\} \\ b &= \mathcal{I}\{c\}. \end{aligned} \tag{1.2}$$

An alternative representation is the *polar* form

$$c = re^{i\theta}, \tag{1.3}$$

where r represents the *magnitude* of c, e is the transcendental number and θ is the *phase* of c. A phase angle of 0 indicates that the number is positive real, and an angle of 180 degrees indicates that it is negative and real.

A *vector* is to be thought of as a point in a multi-dimensional (> 1) space. In this thesis, a typical n-dimensional vector, say $\mathbf{x}$, is written as

$$\mathbf{x} = (x_1, x_2, \ldots x_n), \tag{1.4}$$

where x_i is the i^{th} component (i.e. the amplitude along the i^{th} axis) and n is the number of dimensions.

It is sometimes convenient to group many individual measurements of a particular phenomenon into a *set*. The set of N objects, each denoted by a_i, is written as

$$\{a_i,\ i = 1, 2, \ldots, N\}, \tag{1.5}$$

where a_i is called the i^{th} *element* of the set. The notation $\{a_i\}$ is employed to refer to the whole set as a single entity.

A mathematical function describes the relationship between two variables, e.g.

$$y = f(x), \tag{1.6}$$

where $f(\cdot)$ defines the relationship between the *independent variable* x and the *dependent variable* y. In general, both x and y may be complex-valued (usually shortened to complex) vectors.

1.1.2 Signals

The measurement of a physical phenomenon results in a set of measurement data, which in engineering science is often termed a *signal*. When discussing signals, the measurement process is often ignored and the signal is said to be the phenomenon itself (Oppenheim and Willsky, 1983, Chapter 1). In cases where a particular phenomenon cannot be measured directly (perhaps because it is only a small feature of some other observable phenomenon), it is sometimes convenient to represent that phenomenon by a "hidden" signal when investigating models explaining the phenomenon.For example, in the analysis of sound waves passing through the vocal tract (§2.3.1), it is convenient to consider the behaviour of signals representing the sound waves at various points within the vocal tract, even though these signals cannot be directly measured at those points.

Signals characterise physical phenomena, with information about the phenomena contained in patterns of temporal or spatial variations in the signals (Oppenheim and Willsky, 1983, pp9–11). For example, sounds consist of temporal variations in acoustic pressure, with the identity of any sound determined by a particular pattern of pressure variation. In order to analyse and process signals, it is necessary to represent them as functions of one or more independent variables. For instance, sound can be represented as a function of time, with the acoustic pressure at any instant t represented by some appropriate notation, say $s(t)$. Other independent variables are invoked in the representation of other signals (for example spatial coordinates for the representation of spatially varying phenomena), but in all of the examples included in this thesis, I employ signals which are functions of time.

Signals may be *continuous* functions, as in the example of sound mentioned above (Fig.1.1), or *discrete* functions, defined only at discrete values of the independent variable. Discrete signals may arise from inherently discrete phenomena, such as the weights of each cow in a herd (Fig.1.2), or from successive *samples* of a continuous

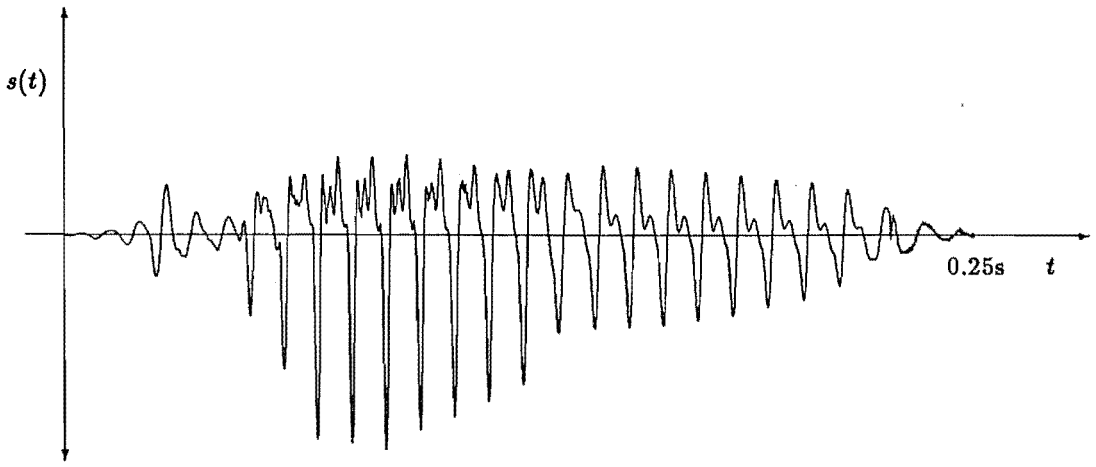


Figure 1.1. Example of a continuous (speech) signal. The vertical axis of the graph represents the acoustic pressure of the sound. Time is represented by the horizontal axis. The signal represents the word "when" spoken by a male speaker.

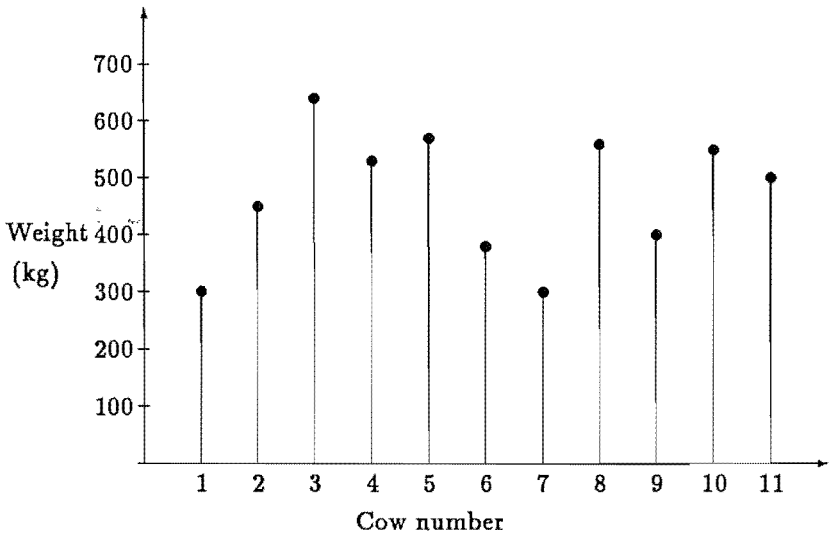


Figure 1.2. Example of a discrete signal, which represents the weights of each of a number of cows. The numbers along the horizontal axis represent the identity of individual cows, while the height of the vertical line at each number indicates the weight of that cow.

process, such as the temperature of a chemical reaction at discrete (say, one second) intervals. Continuous signals must necessarily be sampled if the signals are to be subjected to digital signal processing. The precautions which must be observed to ensure that a sampled signal is a faithful representation of the original continuous signal are detailed in §1.2.5.5. A typical discrete signal is represented mathematically in this thesis by the notation $f[n]$, where n is an integer which refers to the n^{th} occurrence (or sample) of the phenomenon, for which $f[n]$ is the sampled value. For sampled signals, n refers to the instant at which the n^{th} sample is recorded. Samples are usually recorded at regular intervals, with the n^{th} instant occurring at $t = nT$, where T is the *sampling*

interval (see §1.2.5.5).

In order to mathematically analyse the behaviour of real-world signals, they are often *modelled* by simple mathematical functions that behave similarly to the signal. The degree of realism required in the analysis determines how accurately the mathematical function must approximate the real-world signal. For example, the oscillation of a pendulum can be adequately represented for many purposes by a sinusoidal function, i.e.

$$x(t) = A\sin(\frac{2\pi t}{\tau}), \tag{1.7}$$

where $x(t)$ is the (horizontal) displacement of the pendulum at any instant t, A is the maximum displacement, and τ is the time taken for the pendulum to complete one swing back and forth. The properties of sinusoidal functions are well understood, and so mathematical manipulations of them can reveal much information about the behaviour of a pendulum. For a more detailed description of the behaviour of a pendulum, further terms can be added to (1.7) to account for deviations from the simple sinusoidal function (cf. Halliday and Resnick, 1966, p358).

The sinusoidal function occurring in (1.7) is possibly one of the most useful for modelling real-world signals, both because sinusoids are so easy to operate on mathematically, and because signals of this type often arise in real-world situations (Oppenheim and Willsky, 1983, §2.3). It is often useful to re-express a sinusoidal signal as a complex exponential, e.g.

$$z(t) = Ae^{i(2\pi ft+\phi)}, \tag{1.8}$$

where A is the peak amplitude of the signal, $f = 1/T$ is the *frequency* (i.e. the periodicity of the oscillation) and ϕ is the initial phase (defined at $t = 0$). $z(t)$ is complex, as opposed to real-world signals, which are real. However, a complex representation of a signal can facilitate mathematical analysis (such as that presented in §1.2). The "real" signal $x(t)$ is usually associated with the real part of $z(t)$, i.e.

$$x(t) = \mathcal{R}\{z(t)\}. \tag{1.9}$$

A sinusoid is an example of a *periodic* signal, which has the same value at any pair of times separated by the period T, i.e.

$$s(t) = s(t+T). \tag{1.10}$$

The modelling of a complicated signal as a superposition of many different sinusoidal functions is discussed in §1.2. Such a signal remains periodic at a *fundamental frequency* if the frequencies of all the component sinusoids are multiples of the fundamental frequency. The *bandwidth* of a signal is defined as the range of frequencies that its component sinusoids span (see §1.2.5.4 for a more quantitative definition of signal bandwidth).

The factors which define the behaviour of a signal, such as its amplitude A (1.8) or period T (1.10) are here called the *defining characteristics* of that signal. A signal may be classed as *steady-state* if its defining characteristics are invariant over all time, or *transient* when these characteristics change with time (Kreyszig, 1979, p39). Periodic signals are termed steady-state because, even though the actual value of the signal changes with time, it changes in exactly the same way in each period. Hence, if one knows the values of the signal for one period, the behaviour of the signal is known for all time.

Real-world signals are never perfectly steady-state, because they are necessarily of finite duration. However, *segments* of a signal, within which its defining characteristics are (for practical purposes) invariant, may be modelled as being steady-state to

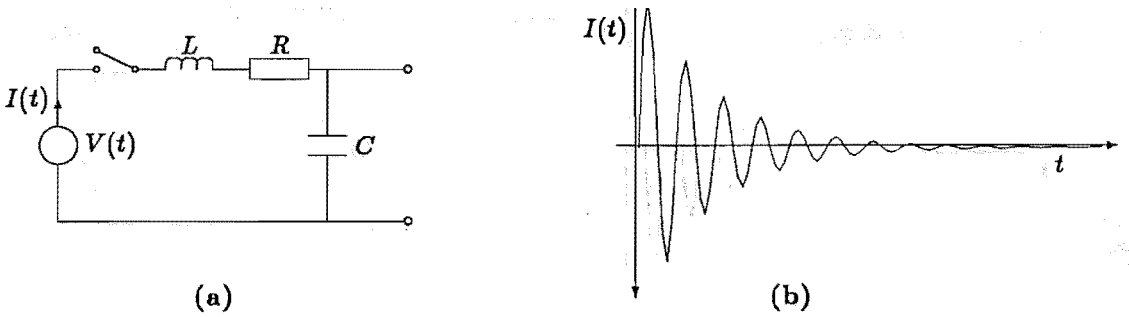


Figure 1.3. Transient signal in an electrical circuit. **a:** *RLC* circuit, **b:** current waveform which results when the switch in *a* is suddenly closed.

facilitate analysis. The transitions between steady-state segments can then be modelled as *transients*. A segment of a signal can be extracted by *windowing*, which involves multiplying the signal by a function that is zero for all times except those within the desired *window* (see §1.3.1.1).

Transients occur when a signal suddenly changes its behaviour. A simple ideal transient is the unit step $u(t)$ defined by

$$\begin{aligned} u(t) &= 1, \quad t \geq 0 \\ &= 0, \quad t < 0, \end{aligned} \tag{1.11}$$

which models a sudden switching on or off of a signal. The exponential function

$$s(t) = e^{\alpha t} \tag{1.12}$$

for which α is real, models a gradually increasing or decreasing phenomenon, with $|\alpha|$ indicating the rate of change. The exponential decay/growth function described by $s(t)$ in (1.12) is often more realistic than the sudden step introduced in (1.11) because real-world phenomena inevitably take finite times to respond to disturbances. Another ideal transient is the unit impulse:

$$\begin{aligned} \delta(t) &= \infty, \quad t = 0; \\ \delta(t) &= 0, \quad t \neq 0; \\ \int_{-\infty}^{\infty} \delta(t)\mathrm{d}t &= 1 \end{aligned} \tag{1.13}$$

which models a spike or short pulse. Such ideal transients and steady-state functions can often be combined in various ways to mathematically model many kinds of simple real-world signals. For example, Fig.1.3*b* shows the signal generated when a voltage is suddenly applied to the electrical circuit shown in Fig.1.3*a*. The current can be approximated by

$$I(t) = u(t)Ae^{-\frac{R}{2L}t}sin(\omega_o t), \tag{1.14}$$

where $u(t)$ represents the switching on of the voltage, $\omega_o = 1/\sqrt{LC}$ is the frequency of oscillation, and R, L, and C are the values of the resistor, inductor, and capacitor respectively.

1.1.3 Systems

A system is thought of in this thesis as any process which transforms an input signal into an output signal. Invoking the terminology utilised previously in this chapter, a system is a mechanism from which a phenomenon arises. For instance, a "hi fi" is a

system that can produce the complicated sound patterns that we know as music (albeit from some pre-recorded source). Systems that are commonly subjected to analysis techniques of the kind described in this thesis always have an input signal, because it is the input which disturbs the system, hence producing the output signal (Oppenheim and Willsky, 1983, §2.5). There may be multiple input and output signals, the forms of which can be very different. For example, an engine may be considered to be a system for which the input signal is the position of the throttle, and the output signal is the speed of rotation of the drive shaft.

The behaviour of a system is described by the system transformation function $L\{\cdot\}$ between the input signal $x(t)$ and the output signal $y(t)$:

$$y(t) = L\{x(t)\}. \tag{1.15}$$

The function $L\{\cdot\}$ is most generally represented as a differential equation. However, direct solution of differential equations is often difficult, in which case various simplifying assumptions about the behaviour of the system are required in order to make the mathematical analysis manageable (Rorabaugh, 1986, §3.2). Such assumptions can lead to more tractable representations of the system function, in terms of an impulse response (described in the next few paragraphs) or transfer function (described in §1.2.5.1), for instance.

A system is said to be *invertible* if distinct inputs lead to distinct outputs (Oppenheim and Willsky, 1983, p39). This property allows the input signal to be deduced from observations of the output signal when the form of $L\{\cdot\}$ is known. A *causal* system is one in which the output signal at any time only depends on present and past values of the input signal (Oppenheim and Willsky, 1983, p40). Both invertibility and causality are satisfied by a *linear system*, which also satisfies the superposition property (Oppenheim and Willsky, 1983, p43):

$$aL\{x_1(t)\} + bL\{x_2(t)\} = L\{ax_1(t) + bx_2(t)\}, \tag{1.16}$$

where a and b are arbitrary scaling constants. A system is said to be *time-invariant* when the only effect of delaying the application of an input signal is to delay the output signal identically (Oppenheim and Willsky, 1983, p42), i.e.

$$L\{x(t - t_d)\} = y(t - t_d), \tag{1.17}$$

where t_d is the delay time.

System analysis is simplified by the assumption of linearity because, once the response of the system is calculated for simple signals (such as the impulse, sinusoid and step functions described in §1.1.2), it can be immediately generalised to other signals by the principle of superposition. Time-invariance ensures that the calculated response of the system remains the same when the signals are applied at different times. Any systems of this kind is conveniently analysed in terms of its *impulse response*, which is the output signal produced by a single impulse (§1.1.2) applied to the input. Because of the linear and time-invariant nature of the system, its response to an arbitrary signal can be derived by considering the input signal to consist of many weighted and time-shifted impulses (Fig.1.4*a*):

$$x(t) = \int_{-\infty}^{\infty} x(t')\delta(t - t')\mathrm{d}t'. \tag{1.18}$$

The output of the system $y(t)$ is then the superposition of many copies of its impulse response, each one shifted by t' and weighted by the amplitude of the input signal $x(t')$ at that instant (Fig.1.4*c*):

$$y(t) = \int_{-\infty}^{\infty} x(t')h(t - t')\mathrm{d}t', \tag{1.19}$$

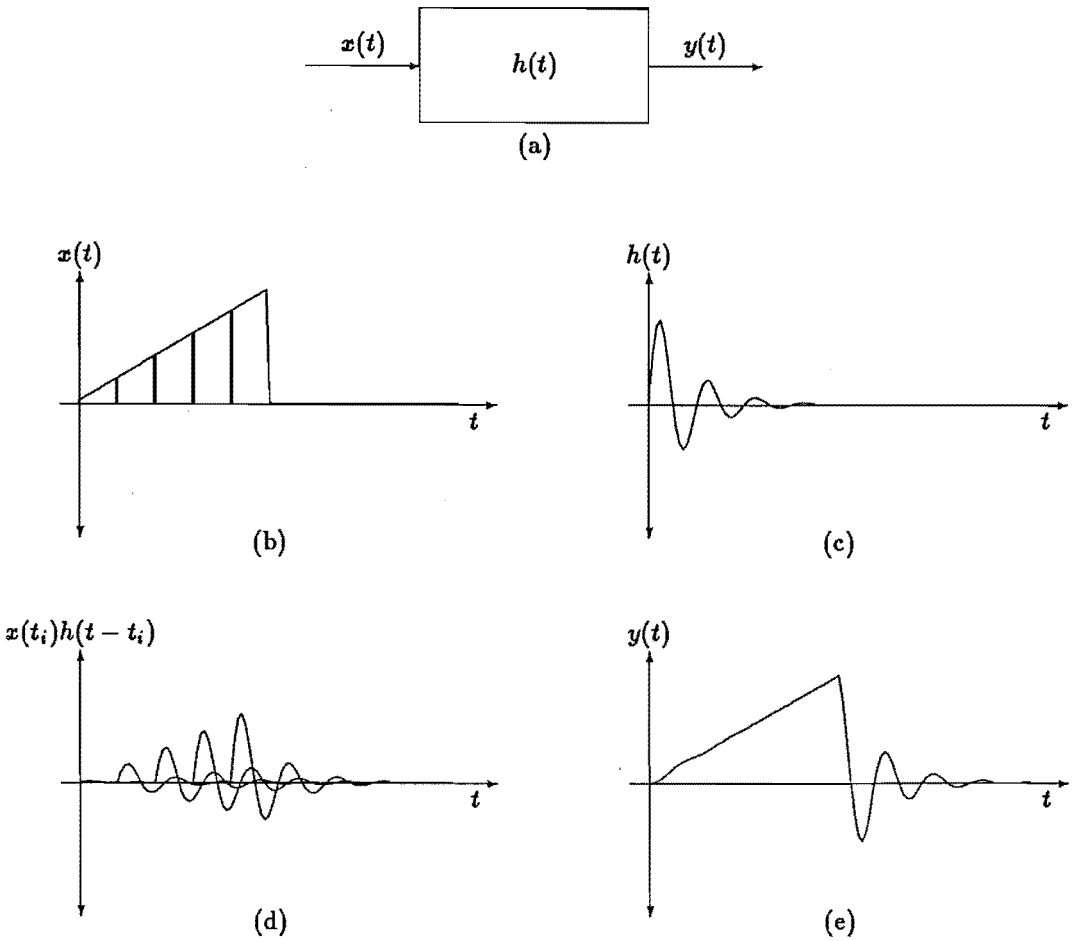


Figure 1.4. Block diagram of a system, illustrating the process of convolution between an input signal and the impulse response of the system. **a:** The input signal, with some of the multitude of impulses which it can be represented by. **b:** The system impulse response "signal". **c:** Some of the copies of the impulse response that have been weighted by the input signal "impulses". **d:** The complete output signal.

which is termed the *convolution integral* (Oppenheim and Willsky, 1983, p90), and is conveniently represented by the shorthand notation

$$y(t) = x(t) \odot h(t) \tag{1.20}$$

where $\odot$ is termed the convolution operator. This important relationship is employed extensively throughout this thesis. Note that convolution is commutative so that $x(t) \odot h(t) = h(t) \odot x(t)$. For linear, time-invariant systems, the quantity $h(t)$, which completely characterises the system transformation function $L\{\cdot\}$, is called the *impulse response* of the system (Oppenheim and Willsky, 1983, p90). Fig.1.4 illustrates the process of convolution between an input signal and the impulse response of a system.

Real-world systems are generally linear for only certain operating conditions. For example, an amplifier may be non-linear for very large input signals, causing "distortion" to the output signal. However, virtually all types of system can be adequately modelled as linear over restricted ranges of input signals. Systems that are not time-invariant, such as the human vocal tract, can often be modelled as time-invariant during

short intervals (§2.3.1.4). Care must then be taken to ensure that the duration of each such interval is consistent with the actual physical process (e.g. for a vocal tract model, short enough that the vocal tract configuration does not change significantly). Notice that the concept of time-invariance for a system is similar to that of steady-state for a signal (see §1.1.2).

1.1.4 Noise

Noise is the term given to any unwanted signals that disturb the transmission and processing of desired signals. Noise is present in any system, and arises from occurrences such as spontaneous random events in electrical circuits, random distortions in transmission media and unwanted interference from other signals (Haykin, 1983, §5.10).

Since noise is by nature random (*stochastic*), it is impossible to predict its exact value at any instant. It is possible, however, to define its properties, and from these derive ways to minimise its influence on the desired signal. The *expected value*, or *mean* of a random signal is the average of all the particular values that can occur. The *variance* is a measure of how much the values are expected to vary from the mean. Another useful descriptor of a random signal is the probability density function (pdf), which specifies the probability, $p(x)\mathrm{d}x$ say, of occurrence of signal values having amplitudes between x and $x+\mathrm{d}x$ (Kreyszig, 1979, §23.7). For example, a random signal with a *uniform* pdf exhibits any value (within a specified range) with equal probability. A random process with a *Gaussian*, or *normal*, pdf is more likely to produce values near to its mean than values further away. The pdf of a Gaussian random process is defined to be (Kreyszig, 1979, §23.10)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1.21}$$

σ^2 is the variance of the Gaussian process, and μ is its mean (§1.3.3). When the defining characteristics of a random signal are time-invariant, the signal is termed *statistically stationary*. Conversely, a random signal is termed *non-stationary* when its defining characteristics vary with time. Further background on the mathematics of stochastic processes and signals can be found in the text by Papoulis (1984).

Noise is said to be *white* when it contains equal power at all frequencies. This is of course physically impossible (because the power would then be infinite), but it is a useful model since the effective bandwidth of a signal or system is often less than the bandwidth of the noise (Haykin, 1983, p279). Signals and systems are conveniently modelled as noise-free, with the effects of noise accommodated by the addition of a noise signal at some appropriate point in the system.

The level of noise in a signal is specified by the *signal-to-noise ratio* (SNR), which is conveniently specified in terms of decibels (dB) (10 times the logarithm to base 10 of the signal-to-noise power ratio). This means that the total SNR of a system constructed from smaller systems can be calculated by adding together the SNRs (in dB) of each component subsystem.

1.1.5 Information

Information is related to knowledge, and, especially in regard to the representation of information by a signal, to knowledge that is not already available to the entity processing the signal (whether that entity be a machine or a human). Intuitively, the amount of information in a signal must depend in some way on the unpredictability of the signal because if future values of a signal are already known then they obviously convey no new information (Cherry, 1978, p170).

For discrete signals, information is defined by the probability of occurrence of each possible *event*, or signal value. For an event with probability of occurrence p_i, the amount of information I_i conveyed by its occurrence is defined as (Shannon, 1948)

$$I_i = -\log_2 p_i \quad \text{bits.} \tag{1.22}$$

A signal must assume some value, and so if there are N possible values, the definition of probability (Woodward, 1953, §1.1) requires that

$$\sum_{i=1}^{N} p_i = 1. \tag{1.23}$$

The *entropy* H of a signal is defined as the average amount of information that is conveyed by each event (Shannon, 1948). So,

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{1.24}$$

which reveals that the maximum amount of information is conveyed by a signal for which each event is equally likely.

When an event X_j occurs as part of a *sequence* $\{X_j\}$ of related events, the information imparted by its occurrence depends on the conditional probability $P(X_j|\mathcal{H}_j)$ that the particular event X_j occurs, given the *history* $\mathcal{H}_j$ of previous events $\{X_{j-1}, X_{j-2}, \ldots, X_{j-N}\}$, where N is the order of the process that generates the events (Hamming, 1980, §5.2). The entropy of such a signal is given by (Hamming, 1980, §6.10)

$$H_N = -\sum P(X_j, \mathcal{H}_j) \log_2 [P(X_j|\mathcal{H}_j)] \tag{1.25}$$

where $P(X_j, \mathcal{H}_j)$ is the probability of the particular sequence $\{X_j, X_{j-1}, \ldots, X_{j-N}\}$ and the summation is taken over all possible sequences. Note that the entropy of a sequence reduces to (1.24) if $P(X_j|\mathcal{H}_j) = P(X_j)$ (i.e. each event is independent).

The entropy is defined in terms of discrete events because the theory of information was spurred largely by the development of digital (telegraph) communications (Hartley, 1928; Shannon, 1948). It can also be applied to continuous signals by sampling the signal appropriately (§1.2.5.5). The number of possible events, or sample values (and hence the probability of each), is enumerated by considering the limit to which each event can be reliably detected, given the SNR of the signal (Shannon, 1948). Using this approach, the information capacity C of a *channel* over which signals can be transmitted is defined, for a channel with a bandwidth B and mean-square SNR S/N, to be

$$C = B \log_2(1 + \frac{S}{N}) \quad \text{bits/sec} \tag{1.26}$$

which is the Hartley-Shannon theorem (Shannon, 1948,1949). It specifies an upper limit on the amount of information that can be conveyed by the channel.

1.2 Signal Representation and transformation

This section introduces the mathematical techniques which are used to transform and analyse signals.

In §1.2.1 I briefly discuss the time and frequency domains, which are commonly invoked when describing different features of a signal. In §1.2.2 I present the Fourier transform and some of its properties. The discrete Fourier transform is described in §1.2.3 and the z-transform in §1.2.4. §1.2.5 discusses some of the implications that

Fourier theory has for the analysis of signals in the time and frequency domains. Finally, §1.2.6 introduces the use of the Hilbert transform for representing signals.

All of the material in this section is presented in university Electrical Engineering courses. Further background can therefore be found in many textbooks, such as those by Oppenheim and Willsky (1983), Haykin (1983), Bracewell (1986), and Bates and McDonnell (1986).

1.2.1 The time and frequency domains

The concept of representing a signal as a function of an independent variable, such as time, is introduced in §1.1. When analysing a signal, or the response of a linear system to a signal, it is sometimes useful (in order to avoid the occurrence of complicated integral expressions like (1.19)), to *transform* the signal into another *domain*, where it becomes a function of some other variable and the integral is replaced by a simpler operation. In addition, certain characteristics of the signal may be more apparent in a different representation (Bracewell, 1986).

A particularly useful way to represent the characteristics of a signal is to transform it into the *frequency domain*. The concept of frequency arises from the study of periodic signals, typified by sinusoidal oscillations. The frequency of a periodic signal is defined as its repetition rate, and the frequency domain representation of a sinusoidal signal is a single component situated at the frequency equal to its repetition rate. The *spectrum* of a signal is the distribution of frequency components from which the signal is constructed.

1.2.2 The Fourier transform

The concept that a periodic signal can be described by a harmonic series was first employed by the ancients in their efforts to predict the occurrence of astronomical events (Neugebauer, 1957). More recently, Euler, Bernoulli and others described the motion of vibrating strings as a superposition of "normal modes", each of which was a sinusoidal oscillation with wavelength equal to an integer divisor of the string length (Oppenheim and Willsky, 1983, pp162–165). Later, Fourier found that any periodic signal could be represented by a series of harmonically related sinusoids (since termed the *trigonometric Fourier series*, Oppenheim and Willsky, 1983, §4.0). Such a Fourier series expansion of a signal $s(t)$ is defined by

$$s(t) = \sum_{n=0}^{\infty} S_n e^{i2\pi n f_o t} \tag{1.27}$$

where f_o is the *fundamental frequency* of the signal, and S_n is the (complex) amplitude of the n^{th} harmonic, which has a frequency of nf_o and a phase at time $t = 0$ of ϕ_n. The coefficients S_n are defined by

$$S_n = \frac{1}{T} \int_{-T/2}^{T/2} s(t) e^{-i2\pi n f_o t} \mathrm{d}t \tag{1.28}$$

where T is the period of the signal (i.e. $f_o = 1/T$). The S_n collectively constitute the *line spectrum* of the periodic signal $s(t)$ (Oppenheim and Willsky, 1983, §4.2).

The concept of representing a signal by sinusoidal components can be extended to aperiodic signals through the use of the *Fourier transform*. This can be simplistically derived from the Fourier series by defining the "period" T of an aperiodic signal as approaching infinity. In this formulation, the fundamental frequency tends towards zero and the individual frequency components of the signal become infinitesimally close

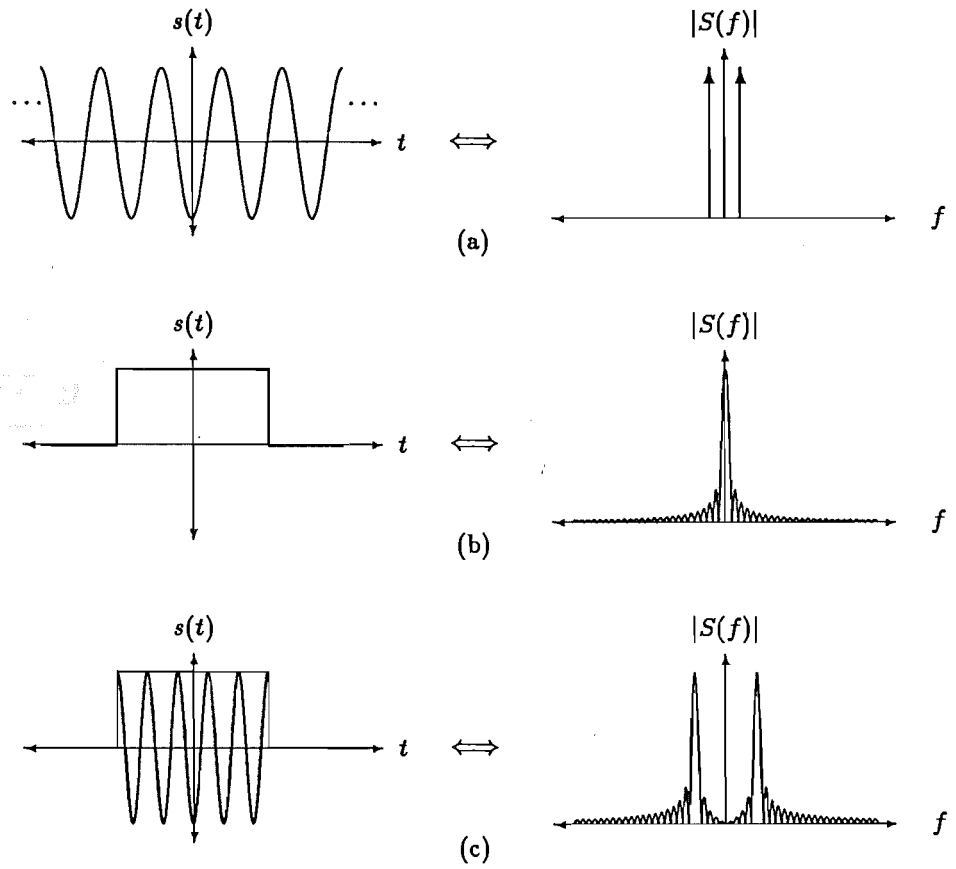


Figure 1.5. Fourier transform pairs. In each case the time domain signal is on the left and its Fourier transform on the right. **a:** Single frequency sinusoid. **b:** Rectangular pulse. **c:** Sinusoidal pulse.

together. The Fourier transform representation of an aperiodic signal is then defined by

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-i2\pi ft} \mathrm{d}t \tag{1.29}$$

and the *inverse Fourier transform* by

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{i2\pi ft} \mathrm{d}f. \tag{1.30}$$

where $S(f)$ and $s(t)$ are termed a Fourier transform pair (Bates and McDonnell, 1986, §6). In this thesis, the notation

$$\begin{aligned} S(f) &= \mathcal{F}\{s(t)\} \\ s(t) &= \mathcal{F}^{-1}\{S(f)\} \end{aligned} \tag{1.31}$$

is used to indicate that $S(f)$ and $s(t)$ are related by the Fourier transform (i.e. they are a Fourier transform pair). Fig.1.5 shows some examples of Fourier transform pairs.

It is often convenient to treat f as a complex variable (Bates and McDonnell, 1986, §13). With this convention, $\mathcal{R}\{f\}$ corresponds to frequency as introduced in

§1.2.1, while $\mathcal{I}\{f\}$ corresponds to the amount of exponential decay ($\mathcal{I}\{f\} < 0$), or growth ($\mathcal{I}\{f\} > 0$). §1.2.5.2 discusses some of the insights into a system's behaviour that can be obtained by examining its response in the complex Fourier domain.

The *Power spectrum*, defined by

$$\Xi(f) = |S(f)|^2 = S(f)S^*(f), \tag{1.32}$$

where $S^*(f)$ represents the complex conjugate of $S(f)$, is a real, non-negative, quantity which indicates the relative intensities of different spectral components of the signal (Bracewell, 1986, pp115–116). The power spectrum, rather than the Fourier transform itself, is commonly employed as a measure of the spectral content of a signal, because the relative amplitudes of the spectral components are unaffected by their phases. However, it is necessary to know the phases if the form of the time domain signal is to be recovered from the spectrum (Bates and McDonnell, 1986). The inverse Fourier transform of the power spectrum is the autocorrelation function of the signal (§1.2.5.1).

1.2.3 The discrete and the fast Fourier transforms

When processing signals it is usually convenient to utilise the power and flexibility of digital computers. These can only operate on discrete numbers, and so a discrete version of the Fourier transform is necessary to perform Fourier processing in practice. It is shown in §1.2.5.5 how a discrete version of a continuous signal can be defined.

The *discrete Fourier transform* (DFT) of a discrete signal $s[n]$ comprising N samples is defined by the relations (Bates and McDonnell, 1986, §12)

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-i\frac{2\pi}{N}nk} \tag{1.33}$$

and

$$s[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] e^{i\frac{2\pi}{N}nk}. \tag{1.34}$$

where the index k identifies the N samples of the Fourier representation of $s[n]$.

The DFT is in effect a trigonometric Fourier series with only N terms. This is a consequence of $s[n]$ and $S[k]$ being sampled, and arises because

$$e^{i\frac{2\pi}{N}n} = e^{i\frac{2\pi}{N}(n+N)}. \tag{1.35}$$

Fig.1.6 illustrates this "periodicity" of the DFT, in that the spectral coefficients repeat after N terms (Fig.1.6). Some of the consequences of this periodicity are dealt with in §1.2.5.5.

The DFT, as expressed by (1.34), requires about N^2 (denoted by $O(N^2)$) multiplications to implement in a simple-minded fashion. However, an elegant algorithm which takes advantage of the common factors in the DFT summation (1.34), called the fast Fourier transform (FFT), enables the DFT to be evaluated with $O(N \log N)$ multiplications (Brigham, 1974). A small drawback of the FFT is that it must be performed on a sequence with an easily factorable length. Most algorithms require that the length be a power of 2, although "mixed-radix" FFT algorithms are available which can operate on sequences of arbitrary length (they are, however, computationally slower, Burrus and Parks, 1985).

1.2.4 The z-transform

The *z-transform* is a discrete transform relationship defined by

$$H(z) = \sum_{n=-\infty}^{+\infty} h[n]z^{-n} \tag{1.36}$$

where $H(z)$ is termed the z-transform of the sequence $h[n]$ and z is a complex variable (Oppenheim and Willsky, 1983, Chapter 10). The variable z is related to the Fourier variable f by

$$z = e^{i2\pi fT}, \tag{1.37}$$

where T is the interval between successive samples. Together with (1.34), this relationship indicates that the unit circle in the z-domain corresponds to the real axis in the complex Fourier domain.

1.2.5 Implications of the Fourier transform

In this section, some of the properties of the Fourier transform and how they pertain to signal processing are discussed. §1.2.5.1 introduces the *convolution theorem*, which can be invoked to reduce the complexity of expressions describing the response of a system to an input signal. §1.2.5.2 describes how a signal or system response function can be described by a factored polynomial. The utility of this representation for modelling systems is also discussed. In §1.2.5.3 I introduce the concept of inconsistency of convolution. Finally, §1.2.5.4 and §1.2.5.5 discuss various practical implications of Fourier-theoretic aspects of time durations and frequency bandwidths (of continuous and sampled signals respectively).

1.2.5.1 The convolution theorem

In the time domain, the response of a linear, time-invariant system to an input signal is described by the integral expression (1.19). However, in the frequency domain, the response is given simply by the multiplication

$$Y(f) = S(f).H(f) \tag{1.38}$$

where $Y(f)$, $S(f)$ and $H(f)$ are the Fourier transforms of $y(t)$, $s(t)$ and $h(t)$ respectively. $H(f)$ is termed the *transfer function* of the system.

The relationship (1.38) illustrates the *convolution theorem*, which states that a convolution between two signals in one domain is equivalent to a multiplication between those signals in the transformed domain (Bates and McDonnell, 1986, §7). Thus, if $S(f) = \mathcal{F}\{s(t)\}$ and $H(f) = \mathcal{F}\{h(t)\}$,

$$S(f).H(f) = \mathcal{F}\{s(t) \odot h(t)\}. \tag{1.39}$$

A consequence of the convolution theorem is that the power spectrum $|S(f)|^2$, defined by (1.32), is the Fourier transform of the autocorrelation function:

$$|S(f)|^2 = \mathcal{F}\{s(t) \odot s^*(-t)\}, \tag{1.40}$$

which is called the *Weiner-Khinchine* or *autocorrelation* theorem.

1.2.5.2 Poles and zeros

The transform (z or Fourier) of a signal can be expressed as a ratio of polynomials of the form (taking the case of the z-transform)

$$H(z) = \frac{\prod_{m=0}^{\infty}(z-\beta_m)}{\prod_{k=0}^{\infty}(z-\alpha_k)} \tag{1.41}$$

where the α_k and β_m are termed the *poles* and *zeros* of $H(z)$ respectively (Oppenheim and Willsky, 1983, §9.4). The positions of the poles and zeros characterise the signal or system response. In the analysis of system response functions, the pole-zero representation is useful because it intuitively conveys information about the resonant or anti-resonant characteristics of the system. A resonance occurs at a frequency at which the input signal tends to be reinforced, while an anti-resonance is said to occur at a frequency at which the input signal is reduced. A resonance introduces a peak into the system transfer function, while an anti-resonance introduces a dip. A pole represents a resonance, with its frequency indicated by its phase (z domain) or real part (Fourier domain) and its damping indicated by its distance from the unit circle (z-domain) or real axis (Fourier domain). Furthermore, the system is stable if all the poles are inside the unit circle (z-domain) or in the lower half plane (Fourier domain) (Oppenheim and Willsky, 1983, §10.7). In the Fourier domain, the zeros always occur on the real axis or in conjugate pairs.

Any system function can be factorised in several ways, although the number of factors required to represent the system can be minimised by choosing a representation that matches the characteristics of the system. The general formulation (1.41) is termed a *pole-zero model*. An *all-pole model* is often used to represent a system that consists of several resonances. The all-pole system transfer function can be written as

$$H(z) = \frac{1}{1-\sum_{k=1}^{P} a_k z^{-k}} \tag{1.42}$$

where P is termed the *order* of the system and the $\{a_k\}$ are the coefficients of the polynomial that characterises $H(z)$. An all-pole model is also termed an *auto-regressive* (AR) or predictor model (Kay and Marple, 1981; Gutowski *et al.*, 1978). This is because the output of such a system can be predicted from previous outputs. By applying a signal $x[n]$ to a system characterised by the $H(z)$ of (1.42), an output

$$y[n] = x[n] + \sum_{k=1}^{P} a_k y[n-k] \tag{1.43}$$

is produced. An all-pole model of this form is often employed in the analysis of speech signals (§3.2.1).

An *all-zero* (often termed a *moving average* or MA) model is represented by the system function

$$H(z) = 1 - \sum_{m=0}^{P} b_m z^{-m} \tag{1.44}$$

where the b_m are the samples of the impulse response of $H(z)$. The number of coefficients required to represent a system's transfer function as an all-zero model is thus

proportional to the duration of its impulse response. By contrast, an all-pole model can characterise a system having an impulse response of infinite duration with only a few coefficients (see §3.2.1).

The use of models such as those described above to represent a system is discussed more fully in §1.3.1.

1.2.5.3 The General inconsistency of convolution

As implied in §1.2.5.2, a signal can be uniquely specified by the zeros of its Fourier transform. Furthermore, only a finite number of zeros are required to represent a discrete signal of finite duration (Bates and McDonnell, 1986, §13). A convolution between two signals, $\lambda[n]$ and $\gamma[n]$, of finite duration, is then given by the signal $z[n]$ that is characterised by the set of zeros Z, for which

$$Z = \Lambda \bigcup \Gamma, \tag{1.45}$$

where Λ and Γ are the sets of zeros of $\lambda[n]$ and $\gamma[n]$ respectively and $\bigcup$ denotes the set union operator (Bates and McDonnell, 1986, §14).

A consequence of (1.45) is that, given an arbitrary signal $x[n]$ of finite duration, which is characterised by a set of zeros X, it is almost never possible to construct a signal $y[n]$ such that $x[n] \odot y[n]$ is exactly equal to $z[n]$. Unless $x[n]$ is constrained so that X is a subset of Z, $z[n]$ and $x[n] \odot y[n]$ are said to be *inconsistent* (Bates and McDonnell, 1986, §14). On the other hand, there are an infinite number of signals $y[n]$ and $w[n]$ for which

$$z[n] = x[n] \odot y[n] + w[n] \tag{1.46}$$

is true, regardless of the form of $x[n]$. The signal $w[n]$ can be thought of as the signal which must be added to $x[n] \odot y[n]$ to move the latter's zeros just enough that they coincide with the zeros of $z[n]$.

1.2.5.4 The bandwidth and time-duration of a signal

If a signal $s(t)$ has a Fourier transform $S(f)$ which is zero for $|f| > W$, it is said to be *band-limited* to W, or to possess a *bandwidth* of W. In addition, a signal that is non-zero only throughout an interval of duration T (e.g. $s(t) = 0$ for $-T/2 < t < T/2$) is said to be *time-limited.* Examination of the Fourier integral (1.29) shows that a signal cannot be both time-limited and of finite bandwidth. However, a little thought about the behaviour of signals and the limitations of systems indicates that neither the bandwidth nor the duration of a real-world signal can be effectively infinite. For example, the sounds made by a person speaking must obviously be time-limited. They must also be band-limited because sound waves of very high frequencies cannot propagate through air. This seeming paradox is resolved by remembering that the mathematical representation of a signal is only an idealised model of the real world. Although the mathematical representation of a time-limited signal has infinite bandwidth, the difference between that and a (real) band-limited signal is too small to make any difference in the real world, as Slepian (1976) makes clear, illustrating his argument with an example which has inspired the one used here. Thus the *effective bandwidth* of a signal is the range of frequencies outside which exist only components of the signal that can be neglected without causing any measurable changes to the signal. Likewise, the *effective duration* is defined as the interval outside which only an insignificant amount of the signal exists.

The question of what the bandwidth of a time-limited signal actually is was first considered in the days of long distance telegraphy (Nyquist, 1924). Later, Gabor (1946)

and Shannon (1949) considered the related question of how much information a channel of a certain bandwidth can carry in a limited time. Shannon expressed this by saying that the "dimension" of a signal of effective duration T and effective bandwidth W is $2WT$ (also see §1.2.5.5), while Gabor derived an "elementary signal" for which the product of bandwidth Δf and time-duration Δt is equal to $1/2$. This type of analysis also led to the quantitative measure of information stated in §1.1.5 and the sampling theorem discussed in §1.2.5.5.

Because the Fourier transform is defined by an integral for all time, it reveals nothing about the positions in time of various components of the signal. This is more than adequate when analysing a steady-state signal (such as the impulse response of a linear, time-invariant system), but it is very inconvenient when the spectral content of the signal varies with time. For example, everyone is familiar with music, and how the frequencies (tones) of the musical notes seem to change with time. This question of a "time-varying frequency" is related to the issues discussed in the previous two paragraphs, and can be resolved in a similar way (cf. Gabor, 1946). A *time-varying* signal can be considered to be time-invariant for short intervals, and hence can be considered as the concatenation of many short segments, each of which is effectively time-limited. The *short-term spectrum* is defined as the Fourier transform of a particular segment of signal, and the *time-varying spectrum* as the (2-dimensional) function obtained by concatenating the short-term spectra for successive segments of signal (§3.3.1). An alternative way of representing the time-varying spectral nature of a signal is by means of its *instantaneous frequency* (cf. Cook and Bernfield, 1967). This is discussed further in the section on the Hilbert transform (§1.2.6).

1.2.5.5 The sampling theorem

When processing signals by computer, the real continuous signal must be represented by a finite number of discrete-valued samples, which must be sufficiently close together that they adequately represent the continuous signal. This *sampling rate* depends upon the bandwidth of the signal.

The *sampling theorem* (Shannon, 1949) states that a signal $s(t)$, of bandwidth W, is uniquely specified by its samples at $s(nT)$, where n is any integer, and $T = 1/2W$ or less (the reciprocal of T, termed the *sampling rate*, or *Nyquist rate*, is often employed in discussions of sampling, Oppenheim and Willsky, 1983, §8.1). The sampling theorem can be derived by considering the samples $s(nT)$ to be a Fourier series expansion of the band-limited spectrum $S(f)$ of $s(t)$. A consequence of this is that the spectrum of a sampled signal is periodic, with a fundamental period $2/T$ (see Fig.1.6). If the (unsampled) signal has non-zero components of frequency greater than $2/T$, the process of sampling forces the energy in these components to lie in the frequency range of 0 to $2/T$, because only these frequencies exist in the sampled signal (see (1.34) in §1.2.3). This "frequency folding" is termed *aliasing*, and implies that the sampled signal does not accurately represent the continuous signal. When real-world signals are to be sampled, such frequency components must be removed by low-pass filtering the signal with an *anti-aliasing filter*. This must in practice have a cutoff frequency of less than half the sampling rate, in order to provide sufficient attenuation to render higher frequencies insignificant.

A continuous signal can be recovered from its samples by low-pass filtering the sampled signal with a cutoff frequency equal to half the sampling rate. An ideal low-pass filter is represented by

$$H(f) = \begin{cases} 0 & \text{for } |f| > W \\ 1 & \text{for } |f| < W \end{cases} \tag{1.47}$$

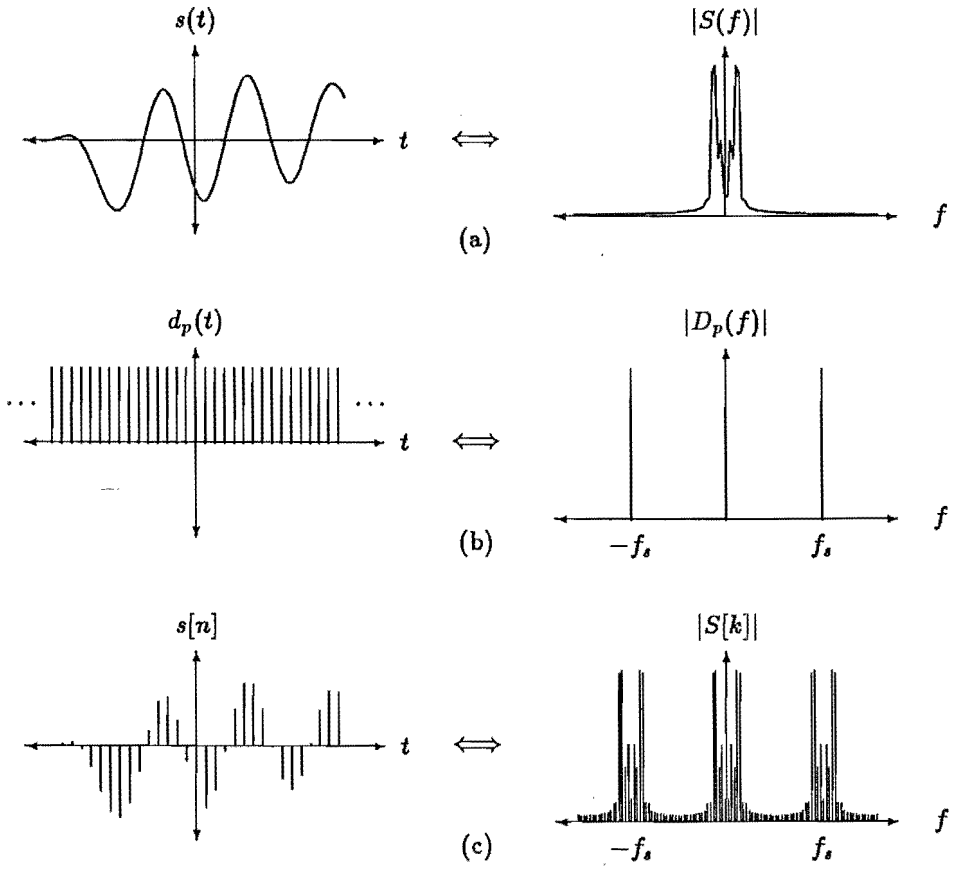


Figure 1.6. An illustration of the sampling process. a: Continuous signal and its spectrum. b: Sampling function. c: Sampled signal. The "spikes" of the sampling function are separated by $T = 1/f_s$ where f_s is the sampling frequency.

where W is the cutoff frequency, and the inverse Fourier transform of this is given by

$$\text{sinc}(t) = \frac{\sin(2\pi W t)}{2\pi W t} \tag{1.48}$$

which is termed the *sinc interpolation* function. $\text{sinc}(t)$ is unity for $t = 0$, and zero for every $t = n/2W$ (i.e. at each of the other sampling instants). Hence when it is convolved onto a sampled signal, it effectively interpolates between each sample and produces a replica of the original continuous signal.

1.2.6 Analytic signals and the Hilbert transform

The Fourier transform of a real signal (1.29) contains both positive and negative frequencies. As already implied in §1.1.2 (in the paragraph containing (1.8)), it is sometimes convenient to employ a (complex) representation of a signal that contains only positive frequencies. Such a representation is the *analytic signal* (Gabor, 1946), which has a spectrum $\Psi(f)$ defined by

$$\Psi(f) = \begin{cases} 0 & , f < 0 \\ 2S(f) & , f \geq 0, \end{cases} \tag{1.49}$$

where $S(f) = \mathcal{F}\{s(t)\}$ and $\psi(t) = \mathcal{F}^{-1}\{\Psi(f)\}$ is the analytic representation of $s(t)$. In the time domain, $\psi(t)$ is defined by

$$\psi(t) = s(t) + i\sigma(t) \tag{1.50}$$

where $\sigma(t)$, which is the *Hilbert transform* of the real signal $s(t)$, is defined by

$$\sigma(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{s(\tau)}{\tau - t}\mathrm{d}\tau. \tag{1.51}$$

The inverse Hilbert transform, giving $s(t)$ in terms of $\sigma(t)$, is

$$s(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\sigma(\tau)}{\tau - t}\mathrm{d}\tau. \tag{1.52}$$

It is sometimes convenient to express $\psi(t)$ in polar form. It is then appropriate to think of $\psi(t)$ as a vector rotating (with time) around the origin of the complex plane, such that its projection onto the real axis at any instant corresponds to the amplitude of $s(t)$ at that instant. This is expressed as

$$\psi(t) = u(t)e^{i2\pi\int_0^t f_i(\tau)\mathrm{d}\tau} \tag{1.53}$$

where

$$u(t) = [s^2(t) + \sigma^2(t)]^2 \tag{1.54}$$

is termed the *envelope* or *modulation* signal, and $f_i(t)$ the *carrier*. $f_i(t)$, which is also termed the *instantaneous frequency* (Haykin, 1983, p279), is the time derivative of the *phase* of $\psi(t)$. The instantaneous frequency is related to the real signal $s(t)$ and its Hilbert transform $\sigma(t)$ by

$$f_i(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left\{\tan^{-1}\left(\frac{\sigma(t)}{s(t)}\right)\right\}. \tag{1.55}$$

The instantaneous frequency can be employed instead of the short-term spectrum (§1.2.5.4) to characterise the spectral content of a time-varying signal (e.g. as traditionally employed in the analysis of frequency modulated signals, cf. Haykin, 1983, Chapter 4). However, it is sometimes more difficult to interpret, since it can attain unbounded positive or negative values (e.g. if there is a phase discontinuity in the signal).

1.3 Signal processing techniques

In this section I describe some of the techniques that are used to extract information from signals. The techniques are discussed with reference to their application to speech and other sounds. However, similar techniques are also employed in applications as diverse as astronomy (Bates and McDonnell, 1986) and geology (cf. Wood and Treitel, 1975).

In §1.3.1 I discuss some of the practical details of estimating the spectral content of a signal. §1.3.2 introduces a few of the techniques by which a signal can be separated into components, each corresponding to the contribution of a particular part of the phenomenon from whence the signal arose. The statistical analysis techniques that are employed in later chapters are introduced in §1.3.3, while §1.3.4 presents several aspects of the practical details of implementing signal processing techniques.

1.3.1 Spectral estimation

As §1.2 describes, the spectrum of a signal characterises its structure by indicating the relative importance of each of the sinusoidal components that make up the signal. The *power spectrum* (§1.2.2) specifies the power in each frequency component of the signal. For example, power spectral analysis of the electricity mains supply indicates the presence and severity of harmonics in the electricity system (cf. Arrillaga *et al.*, 1985).

In the remainder of this thesis I use the term *spectrum* to refer to both the power spectrum and the (Fourier or z) transform of a signal. Only if the meaning is not clear from the context do I explicitly state which type of spectrum is implied. In most cases, the context is evident from the mathematical notation (as introduced in §1.2.2).

The power spectrum of a signal can be simply obtained by applying (1.29) and (1.32) to the signal. In practice, however, the spectrum obtained in this way is often distorted due to the effects of noise and signal truncation. Various techniques have been developed for generating accurate and reliable estimates of the "true" spectral content of noise-corrupted signals of finite durations (Blackman and Tukey, 1958; Kay and Marple, 1981; Roberts and Mullis, 1987). §1.3.1.1 discusses the techniques pertaining to finite signal duration while §1.3.1.2 describes some methods of reducing the effects of noise on the spectral estimate.

1.3.1.1 On the use of windows

In many instances, spectral estimation can only be performed on a (relatively) short segment of the total signal. This may be because the signal is truncated, either due to measurement considerations (e.g. signals representing economic data can only be measured up to the present time) or computational factors (such as the amount of available memory setting a limit on the duration of the signal which is to be analysed). Segmentation is also necessary when one wishes to compute the short-term spectrum (§1.2.5.4) of a signal or invoke the technique called segment averaging (described in §1.3.1.2 below). Truncating a signal in any of these ways introduces a *bias* into the calculated spectrum. This section describes the origin of the bias and the means by which it can be reduced. Note that such considerations do not apply to signals that are naturally time-limited, implying that their spectra can be computed for their entire durations (as is the case for the sonar clicks which are analysed in Chapter 7). However, "leakage" may still occur if the signal contains spectral components with periods that are not exact integer fractions of the analysis duration (as described in the paragraphs that follow).

Fourier analysis of a finite duration *analysis segment* of a signal reduces to constructing a Fourier series with fundamental period equal to the duration of the segment (§1.2.2). If the signal contains spectral components that are not exact harmonics of this fundamental, they are not accurately represented in the Fourier series. Instead, the energy in such components *leaks* into the other spectral components (Harris, 1978). Leakage can be understood by noting that multiplying a signal $s(t)$ by a window $w(t)$ (which is what extracting a segment amounts to) is equivalent to convolving the signal's spectrum $S(f)$ with the Fourier transform of the window $W(f)$. Hence the estimated power spectrum $\Xi(f)$ is given by

$$\begin{aligned} \Xi(f) &= \mathcal{F}\{s(t)w(t-\lambda)\} \\ &= |S(f) \odot W(f)|^2 \end{aligned} \tag{1.56}$$

where λ is the instant at which the window is applied to $s(t)$ (Blackman and Tukey,

1958). If the window $w(t)$ is a simple *rectangular window*,

$$\begin{aligned} w(t) &= 1, \quad 0 < t < T_n \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{1.57}$$

where T_n is the duration of the window, then its equivalent spectral smoothing window is (Blackman and Tukey, 1958, p95)

$$\begin{aligned} W(f) &= \frac{\sin(2\pi f T_n/2)}{2\pi f} \\ &= T_n \operatorname{sinc}(f), \end{aligned} \tag{1.58}$$

which exhibits large *side-lobes* at frequencies that do not satisfy

$$f = \frac{k}{T_n} \tag{1.59}$$

where k is an integer. These side-lobes are the source of the leakage identified in the first two sentences in this paragraph. Note that the above reasoning is valid whether the continuous or discrete Fourier transform is employed (Blackman and Tukey, 1958).

Several window functions have been proposed that have lower sidelobes than the rectangular window discussed above (Harris, 1978). Particularly simple windows are described by the trigonometric series

$$\begin{aligned} w(t) &= a_0 + \sum_{j=1}^{J} a_j \cos(j\frac{2\pi t}{T_n}) \qquad 0 < t < T_n \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{aligned} \tag{1.60}$$

where the coefficients a_j are chosen to minimise the side-lobe level (Harris, 1978). Note that the integer J is rarely taken to be greater than 4. Table 1.1 lists values of the coefficients characterising several common windows, together with their maximum side-lobe levels with respect to the main lobe. The side-lobe level indicates to what extent the energy from non-harmonic (as described in the previous paragraph) spectral components spreads across the computed spectrum. Windows that have lower side-lobe levels necessarily have a wider main lobe (see the discussion in the next paragraph), which means that the spectral resolution is somewhat reduced when they are employed. However, this drawback is far outweighed by the advantages of limiting leakage to only a few spectral coefficients around each non-harmonic frequency component.

The resolution Δf with which spectral features can be ascertained is limited to (cf. §1.2.5.4)

$$\Delta f \simeq \frac{1}{2T_n} \tag{1.61}$$

where T_n is the duration of the signal which is to be analysed. The use of a non-rectangular window effectively reduces the duration T_n by tapering the signal values near the ends of the segment. The reduction in resolution can also be appreciated by recalling (1.56), which states that windowing is equivalent to smoothing the spectral coefficients.

The windows described by (1.60) can also be applied in the frequency domain with equivalent effect. A DFT of a rectangular-windowed signal can be windowed by convolving with a spectral smoothing window consisting of only $2J + 1$ terms:

$$W[f] = a_0\delta[f] + \sum_{j=1}^{J} \frac{a_j}{2} \left[\delta[f - \frac{j}{T_n}] + \delta[f + \frac{j}{T_n}]\right]. \tag{1.62}$$

Window name	side-lobe attenuation (dB)	Coefficients			
		a_0	a_1	a_2	a_3
Rectangular	13	1	0	0	0
Hanning	31	0.5	-0.5	0	0
Hamming	43	0.54	-0.46	0	0
Blackman	58	0.42	-0.50	0.08	0
3-term Blackman-Harris	67	0.42323	-0.49755	0.07922	0
4-term Blackman-Harris	92	0.35875	-0.48829	0.14128	-0.01168

Table 1.1. Table of common spectral weighting windows, showing their coefficient values (according to (1.60)) and resulting side-lobe suppression level (after Harris, 1978). Note that the signs of the coefficients alternate if the window is constructed in the interval $0 < t < T_n$, but are all positive if the window occupies the interval $-T_n/2 < t < T_n/2$.

Such a smoothing calculation is a feasible alternative to time domain multiplication in two types of situation. One occurs when a Hanning window is employed. (1.62) can then be implemented by means of additions only, because the coefficients are divisions by two or four, which can be digitally implemented as simple binary shifts. The other situation in which spectral smoothing is appropriate is when the spectral smoothing technique described in §1.3.1.2 is employed to reduce the effect of noise on the spectrum.

1.3.1.2 Reducing the effect of noise

If the signal is corrupted by noise, the spectral coefficients obtained by Fourier transforming the signal are also corrupted by noise. In addition, the variance of each coefficient fails to decrease as the analysis duration is increased (cf. Oppenheim and Schafer, 1975, §11.3; Schwartz and Shaw, 1975, Chapter 4). Several methods for reducing the variance of each spectral component have been proposed (cf. Blackman and Tukey, 1958).

Perhaps the most obvious way of reducing the noise in the spectral estimate is to calculate the power spectra of many short segments of the signal and then average them (Schwartz and Shaw, 1975, Chapter 4). When the noise on each spectral estimate is independent, the average spectrum tends to be closer to the true spectrum than the individual spectral estimates are (Fig.1.7). Using this technique, the spectrum $\Xi[f]$ of a signal $s[n]$ is estimated by

$$\hat{\Xi}[f] = \frac{1}{M} \sum_{m=1}^{M} [\mathcal{F}\{s[n]w_s[n - n_m]\}]^2 , \tag{1.63}$$

where $w_s[n]$ is a *signal window* that, when located at the instant n_m, delineates the m^{th} of M segments. The choice of window is subject to the considerations discussed in §1.3.1.1.

Another way of reducing the variance of the spectral estimate is to compute the spectrum for a longer segment of the signal and then smooth several adjacent spectral coefficients together (Schwartz and Shaw, 1975). If the noise on adjacent coefficients is independent, the smoothed spectrum is a better estimate of the "true" spectrum (Fig.1.8). Smoothing of the spectrum can be expressed as a convolution

$$\hat{\Xi}[f] = \sum_{m=-L}^{L} W[f - m]|S[f]|^2, \tag{1.64}$$

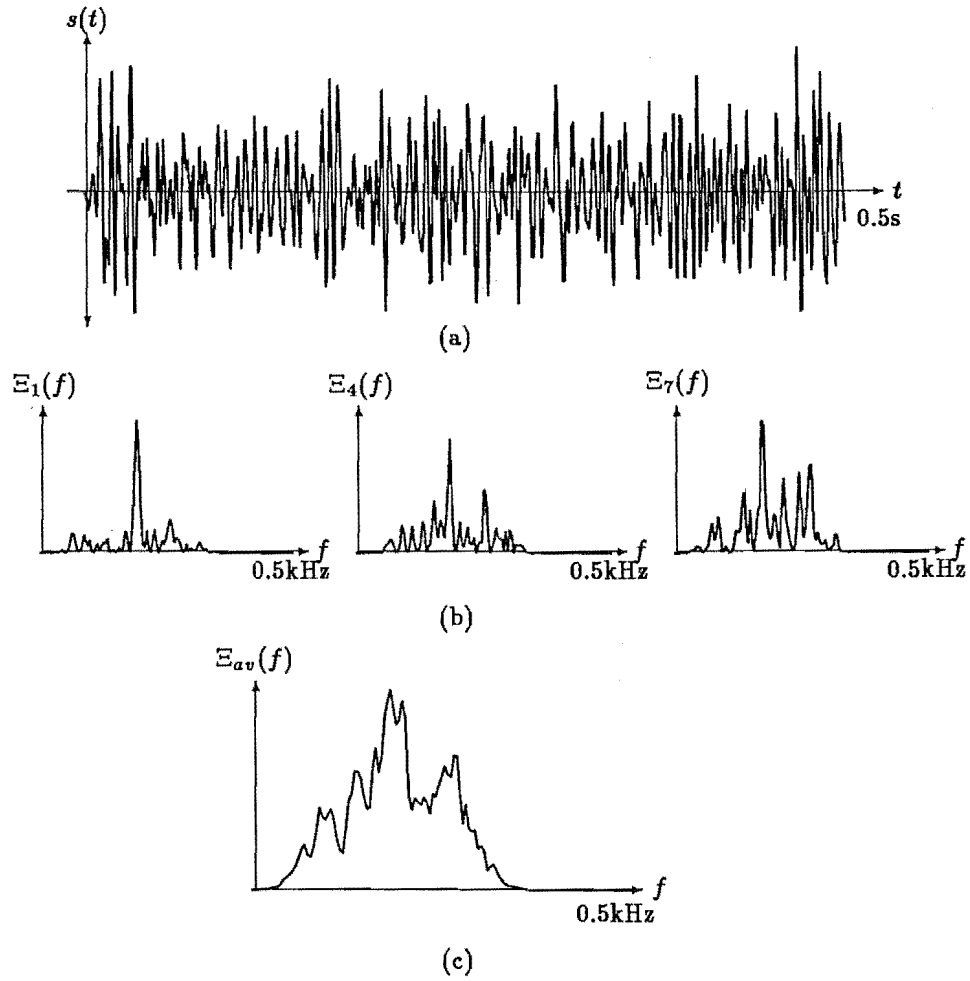


Figure 1.7. Spectral estimation by averaging the power spectra of several short segments. **a:** Part of the original signal. **b:** The spectral estimates of three 256ms long segments of the signal shown in *a*. Each segment was multiplied by a 3-term Blackman-Harris window before performing the FFT. Adjacent segments were overlapped, with a separation of 64ms between each one. **c:** The final spectral estimate computed by averaging all (28) short-term spectra.

where $S[f] = \mathcal{F}\{s[n]\}$ and $W[m]$ is the smoothing window of extent $2L + 1$ samples (see (1.62) in §1.3.1.1). Application of the the autocorrelation and convolution theorems (§1.2.5.1) to (1.64) indicates that smoothing adjacent power-spectral coefficients is equivalent to multiplying the autocorrelation of the signal with a window. $\tilde{\Xi}[f]$ is then given by

$$\tilde{\Xi}[f] = \mathcal{F}\{R[k]w_a[k]\}, \tag{1.65}$$

where

$$R[k] = \sum_{n=-\infty}^{\infty} s[n]s[n-k] \tag{1.66}$$

is the autocorrelation of $s[n]$ and $w_a[k]$ is the *lag window* (Blackman and Tukey, 1958). The presence of $w_a[k]$ ensures that only those values of $R[k]$ for small $|k|$ are used to calculate $\tilde{\Xi}[f]$. This is necessary because, for signals $s[n]$ of finite extent, only a few samples of $s[n]$ appear in the summation (1.66) when $|k|$ is an appreciable fraction of the signal's extent. Hence the averaging, implicit in (1.66), is less effective for larger

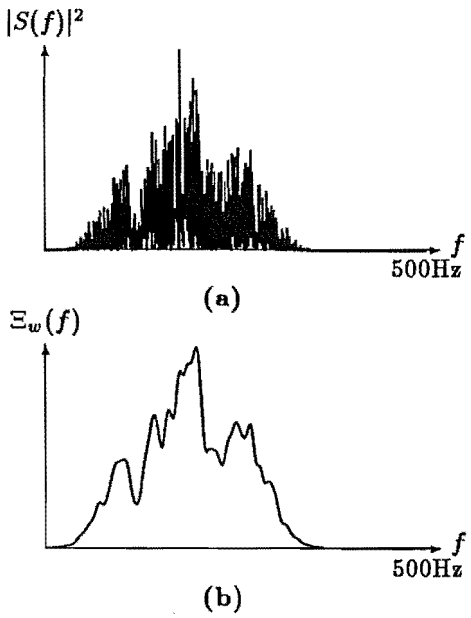


Figure 1.8. Spectral estimation by windowing the autocorrelation. The original signal is the same as in Fig.1.7*a*. **a:** The power spectrum computed from the entire autocorrelation of the signal (which is 2 s in duration). **b:** The smoothed spectrum obtained by applying a 25.6ms duration, 3-term, Blackman-Harris window (see Table 1.1).

values of $|k|$, resulting in higher levels of noise on those components of $R[k]$ (Schwartz and Shaw, 1975, §4.3).

The spectral smoothing and segment averaging methods are equivalent when the durations of $w_a[k]$ and $w_s[n]$ are the same and $w_a[k]$ is the autocorrelation of $w_s[n]$. This can be demonstrated by comparing the expected values of the two expressions, (1.63) and (1.65) respectively, for $\tilde{\Xi}[f]$. The mathematical details of this comparison can be found in several texts, such as the one by Blackman and Tukey (1958, pp92–95).

As mentioned in §1.3.1.1, the spectral resolution is inversely proportional to the duration of the signal which is being analysed. When the spectral smoothing or segment averaging techniques described above are invoked, the effective duration (for the purposes of determining the spectral resolution) is less than the total duration of the signal. Hence the use of these techniques reduces the maximum attainable spectral resolution. For the segment averaging technique, the effective duration is equal to the duration of $w_s[n]$, while for the spectral smoothing technique, it is equal to the duration of $w_a[k]$. Techniques for improving the resolution of a spectral estimate, by taking advantage of any available *a priori* information concerning the signal, are described in §1.3.1.4.

1.3.1.3 Practical considerations for Fourier methods

Whether the spectral smoothing is performed by means of (1.64) or (1.65) depends on computational considerations. Before the FFT algorithm was available, it was usual to employ (1.65), calculating only those values of $R[k]$ embraced by the window $w_a[k]$ (Blackman and Tukey, 1958).

With the advent of the FFT (Brigham, 1974; Burrus and Parks, 1985), it is more efficient to compute $S[f]$ directly from $s[n]$ and then employ (1.64) to estimate the power spectrum (Bingham *et al.*, 1967; Yuen, 1979, Chapter 5). Because of the

implicit autocorrelation that occurs when the power spectrum is obtained by means of (1.63) or (1.64), it is important that the signal $s[n]$ be *zero-extended* to at least twice its duration before computing the FFT (Yuen, 1979, p98). Zero-extension (often called zero-packing in image processing contexts) is a technique of concatenating a number of zero-valued samples on to the ends of the signal. It effectively increases the resolution of the resulting Fourier transform by interpolating between the spectral coefficients that would have been obtained if no zeros were added. Zero-extending is necessary so that the autocorrelation (which is twice the extent of the signal) does not "alias" and so introduce errors into the spectral coefficients.

1.3.1.4 Parametric methods

The resolution limit (1.61) is inappropriate if the duration of the signal that is available for analysis is shorter than that required for the desired resolution. This may, for instance, occur when the signal is truncated (as described in the first paragraph of §1.3.1.1). The resolution may also be insufficient when a signal must be segmented in order to compute the "time-varying spectrum" (see §1.2.5.4). Remember that the time-varying spectrum makes best sense for a signal that can be naturally and usefully divided into many short segments. The spectrum for each segment is computed, and the ensemble of "short-term spectra" then serves to describe the variation of the signal with time. Methods of constructing a time-varying spectrum from such a signal are described further in §3.3.1.

One way to obtain higher resolution of spectral features than what is implied by the signal's duration, is to assume *a priori* that the signal fits a certain model, with only the parameters of the model to be found. For example, a certain signal might be assumed to contain several sinusoidal components of arbitrary amplitude, frequency and phase, together with white noise. Such an assumption implies that a minimum error optimisation scheme is able to estimate the parameters of the sinusoidal components with much greater precision than if no assumptions were made about the signal (Kay and Marple, 1981). However, if the signal contains components that are not accounted for in the model, incorrect estimates are unavoidably produced. Speech sounds are often modelled by a finite number of (complex) poles (§3.2). Each pole characterises the frequency and bandwidth of one of the resonances of the speech production mechanism (§2.3.1). Such a model appears to satisfactorily account for the important characteristics of most speech signals (§2.3.1.3). The techniques used to estimate the model parameters from the speech waveform are described in detail in §3.2.2.

1.3.2 Deconvolution techniques

Some signals are usefully modelled as a convolution between two (or more) components (§1.1.3). For example, a signal that has been (linearly) distorted by some recording apparatus can be expressed as the convolution of a *true signal* and a *blurring signal*, where the latter is the impulse response of the recording apparatus (Bates and McDonnell, 1986; Davey, 1989). If the true signal is represented by $y(t)$ and the blurring signal by $h(t)$ then the recorded signal $r(t)$ is given by

$$r(t) = y(t) \odot h(t) \tag{1.67}$$

in the time domain, and

$$R(f) = Y(f).H(f) \tag{1.68}$$

in the frequency domain, where (as usual) the functions identified by uppercase and lowercase letters are Fourier transform pairs. The process of separating the two components of the recorded signal is called *deconvolution* or *inverse filtering* (Bates and McDonnell, 1986, Chapter III).

Rearrangement of (1.68) results in

$$Y(f) = \frac{R(f)}{H(f)} \tag{1.69}$$

for the case when $H(f)$ is known. Unfortunately, $|H(f)|$ may be small or zero for certain frequencies, which causes errors in the estimate of $Y(f)$. This difficulty is especially acute when either $R(f)$ is corrupted with noise or $H(f)$ is only an estimate, and hence subject to error. A better procedure for estimating $Y(f)$ is called *Wiener filtering* (Wiener, 1949), in which (1.69) is replaced by (Bates *et al.*, 1982b)

$$Y(f) = \frac{R(f)H^*(f)}{|H(f)|^2 + \Phi(f)}, \tag{1.70}$$

where $\Phi(f)$ is chosen to equal the estimated noise-to-signal ratio. In many cases $\Phi(f)$ is set to a constant, due to practical difficulties in estimating the frequency dependence of the noise (Bates *et al.*, 1982b). Note that (1.70) reduces to (1.69) if there is no noise.

Deconvolution can also be performed in the time domain (Bates *et al.*, 1982c). One method of doing this is described at length in Chapter 5 of this thesis. Time domain deconvolution is useful in situations where $|H(f)|$ contains zeros (Högbom, 1974) or when $H(f)$ is time variant (Bates *et al.*, 1982c).

If neither $Y(f)$ nor $H(f)$ is known then (1.70) cannot be applied. The process of separating the two components, when neither is known *a priori*, is called *blind deconvolution* (Stockham *et al.*, 1975; Lane and Bates, 1987). Blind deconvolution techniques can only be applied to signals having two components with properties different enough that they can be separated. The particular technique that is appropriate for any particular signal depends on the way in which the components differ. One useful approach, which is actually equivalent to the original instance of blind deconvolution (Stockham *et al.*, 1975), involves considering the signal to be comprised of many segments, each of which is a convolution of a component that is the same for all segments and one that varies appreciably between segments. The segments can then be suitably averaged to isolate the invariant component. The averaging can be performed in the time domain, as exemplified by the shift-and-add technique described in Chapter 4, or in the frequency domain, as described by Stockham *et al.* (1975).

A second approach, called *homomorphic deconvolution*, can be used to separate two signals when one has a smooth spectrum while the other has a spectrum containing harmonic components (Oppenheim and Schafer, 1968). The use of homomorphic deconvolution in speech processing is described in §3.3.2.

1.3.3 Statistical analysis techniques

Measured data are almost always subject to some random fluctuation (§1.1.4). In addition, many interesting signals are "stochastic" in nature. For example, the signal that describes the variation of temperature from day to day contains patterns relating to seasonal weather changes, but these may be partially obscured due to the random and unpredictable fluctuations inherent in the weather. The random variation inherent in both of these situations can be modelled by a *random signal*, which is conveniently characterised with the aid of *descriptive statistics* (Kreyszig, 1970). Such statistics enable particular characteristics of the signal's behaviour to be predicted (although

they do not allow its value at any particular time to be predicted, as exemplified by the difficulties facing weather forecasters!). For example, the *mean*, or *expected value* describes the average of all the signal values. For a signal comprising N measurements $\{x_i\}$, the mean (denoted by $\bar{x}$ or $\langle x \rangle_i$) is defined by (Kreyszig, 1970, §3.2)

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1.71}$$

The *variance* σ^2, which characterises the spread of signal values from their mean value, is defined by (Kreyszig, 1970)

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}. \tag{1.72}$$

For a continuous signal, having a pdf (§1.1.4) $p(x)$, the mean and variance are given by (Woodward, 1953)

$$\bar{x} = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x \tag{1.73}$$

and

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x) \mathrm{d}x - [\bar{x}]^2 \tag{1.74}$$

respectively. The *standard deviation*, which is the square root of the variance, is often more useful than the variance as a descriptor of signal variation, because it has the same dimensionality as the signal itself.

The mean and variance can be misleading descriptors of a random process, because they are strongly affected by the shape of the pdf (Kreyszig, 1970, Chapter 6). If the pdf is asymmetrical, other descriptors are sometimes more illuminating. The *median* is defined as the value that occurs half way through an ordered list of the signal values (for a discrete signal). For a continuous signal with pdf $p(x)$, the median is a root of the equation

$$\int_{-\infty}^{x} x' p(x') \mathrm{d}x' = \frac{1}{2}. \tag{1.75}$$

The *mode* is the signal value that occurs most frequently (equivalently, a maxima of $p(x)$).

The statistical descriptors introduced in the previous two paragraphs are often obtained from a subset or *statistical sample* of the entire set of signal values. The descriptors obtained from several subsets are themselves subject to statistical variation, because each subset contains different signal values (Snedecor and Cochran, 1980, §5.1). In order to determine if one set of measurements or signal values belongs to the same *stochastic process* as another set, the statistical descriptors of each set are subjected to *hypothesis tests*. A hypothesis test is an assessment of whether the differences between two sets of measurements are *significant* (Kreyszig, 1970, Chapter 11). Various types of tests are employed, depending upon the particular pdf possessed by the signal. Details on these can be found in textbooks such as the one by Kreyszig (1970).

Hypothesis testing is employed extensively in many areas of scientific endeavour. Measurements are obtained under differing experimental conditions, and are statistically analysed in order to determine whether the conditions significantly affect the phenomenon being observed (Snedecor and Cochran, 1980, Chapter 5). Hypothesis testing is also the basis for the techniques, used in the communication and echo-location fields, for detecting signals that are corrupted with noise (Woodward, 1953).

In many situations, several different quantities may be measured that together characterise a particular phenomenon. For example, the weather conditions in a particular locality could be characterised by recording, at a certain time each day, a set of measurements. Each set, which can be represented as a *feature vector* in multi-dimensional space, may consist of the temperature, air pressure, rainfall and humidity. In order to analyse the variation of such a signal, *multi-variate* statistical techniques are required (Cooley and Lohnes, 1971). Some of these techniques are descriptive, in the sense that they attempt to describe the distribution of measurement vectors within the multi-dimensional *feature space*. One such technique is *factor analysis*, which attempts to reduce the dimensionality of the feature space by appropriately combining features that are highly correlated with each other (Cooley and Lohnes, 1971). Reducing the dimensionality also means that the distribution can be more easily visualised (Tukey, 1983).

Multi-variate hypothesis testing techniques are also available. They can be invoked to assess whether the measurement vectors can be separated into different classes or *clusters* (Krishnaiah and Kanal, 1982). Techniques of this type are employed for speech recognition (§3.6.1), vector quantisation of speech signals (§3.5.1.2), and the analysis of the distribution of measurement data in the biological and social sciences (cf. Chapter 7).

1.3.4 The "mechanics" of signal processing

Because of the power and speed of digital computers, they are employed in nearly all practical applications of signal processing techniques. Several aspects of the "mechanics" of processing signals on computers deserve mention here. A block diagram of the steps involved in the processing of signals is shown in Fig.1.9.

Signals appear in many forms (for example sound pressure waveforms, temperature variations or visual images), but in order to process them with a computer, they must be transformed into sets of discrete numbers. Usually, the physical phenomena are first transformed into electrical signals. The devices which perform this transformation are termed *transducers*, which for sound waveforms are commonly called microphones. Because of the great variety of sounds that occur, many types of microphones are available, each suited to a particular range of sound intensities and frequencies. Hence the microphone employed for a particular application should be chosen so that its specified performance matches the characteristics of the sounds that are expected.

A signal that is a continuous function (of time) must be *sampled* before it can be digitally processed. The mathematical aspects of sampling are considered in §1.2.5.5. The *Nyquist rate* specifies the minimum rate that the signal can be sampled at to preserve all the information in the signal. However, practical anti-aliasing filters (§1.2.5.5) do not perfectly attenuate frequency components that are only slightly above the filter's cutoff frequency, and so it is sometimes advantageous to sample at greater than the Nyquist rate.

Sampled signals are stored in the computer as arrays of (binary) numbers. The accuracy of the digital representation depends on the number of binary digits (bits) that are employed for each number. For most signals, 16 bits are more than adequate, and most commercially available analogue-to-digital converters produce at least 12 bits. In practical situations, when the amount of storage is required to be minimised, 8 bits are often employed (§3.5.1). The amount of computer storage that is required for sampled signals depends upon the sampling rate as well as the number of bits used for each sample. For example, signals that are sampled at a rate of 20kHz, with 16 bits (2 bytes) allocated for each sample, occupy 40kBytes of storage per second. At this rate,

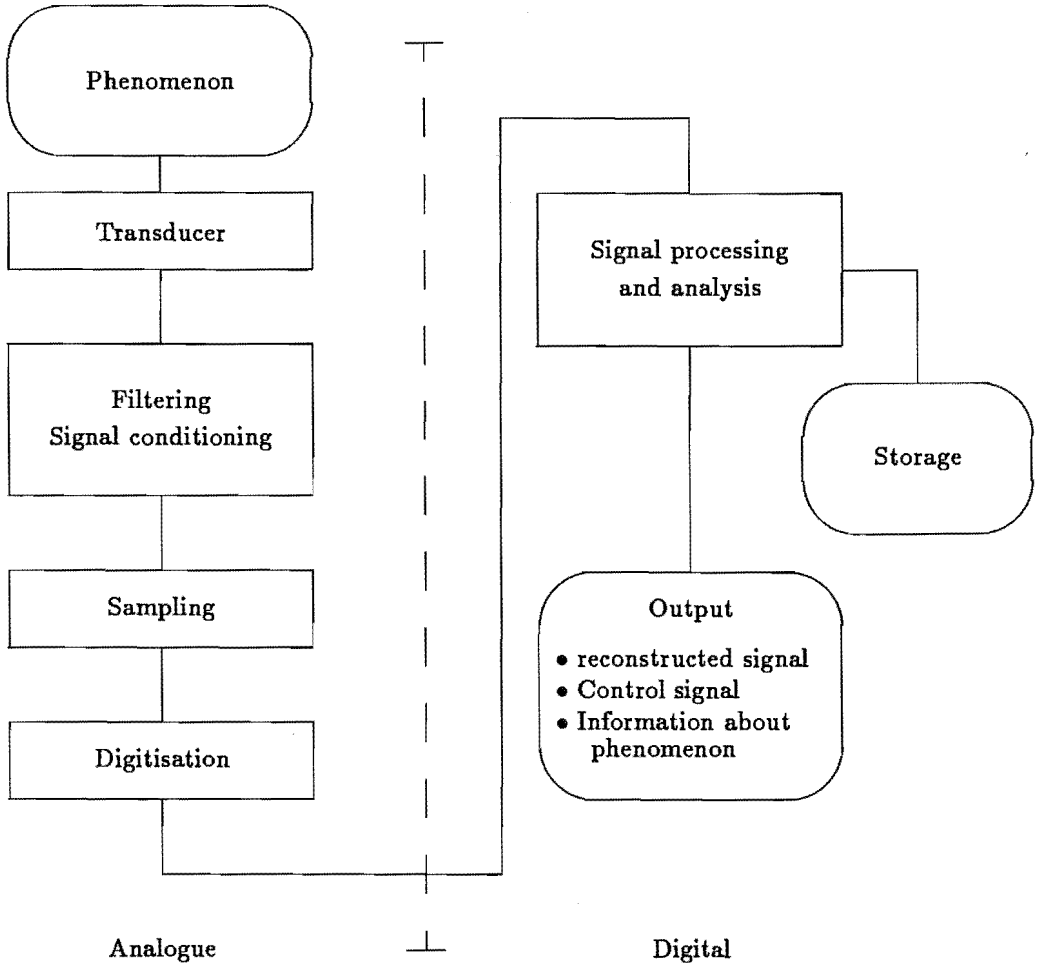


Figure 1.9. Diagram of the steps involved in processing a real-world signal on a digital computer.

a hard disk on a personal computer of 40MBytes storage capacity (typical with present technology) is able to store signals having a total duration of about 15 minutes.

The speed with which the available computational facilities can process the signals is an important consideration in signal processing engineering. This is especially pertinent in *real-time* applications, where the processing must be completed in the same or less amount of time than the duration of the signal. Special purpose "digital signal processing" (DSP) computers are used for some real-time applications such as speech encoding or recognition (cf. Allen, 1975; Burrus and Parks, 1985; Watson *et al.*, 1988). DSPs often have limited programming features, and so for the development of processing algorithms, or the analysis of complicated signal models, the use of large computers and high-level languages is necessary.

Signal processing algorithms can be programmed in any language that allows mathematical operations to be performed on numbers. Much of the programming of the algorithms presented in this thesis was performed in the FORTRAN (Kaufman, 1978) and MODULA-2 languages (Wirth, 1983). For the preliminary development of the algorithms, a specialised signal processing computer language was used. This language, called "SIGPROC", was developed by Brieseman *et al.* (1989) for the purpose of facilitating their speech processing research. It allows signals to be manipulated and processed at a high level of abstraction. In addition, its interactive nature allows operations to be performed by simple commands, with their effects being immediately assessable from a graphical representation of the signal on a computer terminal.

SIGPROC also provides a framework within which more specialised signal processing operations (such as those described in succeeding chapters) can be easily incorporated.

1.4 Introduction to sounds

In this section I give a brief introduction to acoustics, in §1.4.1. I then introduce, in §1.4.2, the types of biological sounds that are the subject of the remainder of this thesis. §1.4.3 describes the speech utterances that are employed in the chapters concerned with speech processing techniques. The procedures by which these speech sounds were recorded is described in §1.4.4.

1.4.1 Acoustics

This thesis is concerned with techniques for the measurement and analysis of certain types of (biological) sounds (§1.4.2). Sound is a physical phenomenon consisting of vibrational waves in a medium. These are generally known as acoustic waves, and the science of studying them is known as acoustics. For the purposes of this thesis, I only consider vibrations that are relevant to the particular biological context (i.e. sounds in air that are necessary for speech communication, vibrations in air and body tissue generated by coughs, and sounds in water produced by dolphins).

A sound is generated by some energy source which disturbs a medium (e.g. the vibration of a guitar string when it is plucked, or the turbulence caused by a strong wind rushing past a building). Hence the nature of the sound in the medium is related to the nature of the disturbance. The disturbance produces waves of compression and rarefaction which travel through the medium. The speed at which the waves travel is determined by the acoustic properties of the medium (and is about 330m/s in air and 1500m/s in water, Kinsler *et al.*, 1982, §5.6,15.2). Hence, when a sound wave meets a boundary between media of different acoustic properties, part of it is reflected and part refracted (Kinsler *et al.*, 1982, Chapter 6). This means that the nature of the sound is modified by its passage through the media.

More details about sounds, their mathematical representation, and the properties of acoustic vibrations in general, can be found in any textbook on acoustics, such as those by Kinsler *et al.* (1982) or Morse and Ingard (1968).

1.4.2 Biological sounds

As mentioned in the previous section, sounds are induced in a medium when something causes a disturbance to the medium. In the context of the biological world, sounds are produced when part of an animal moves. Sounds can be classified as intentional and non-intentional, depending on whether or not the primary intention of the movement is to generate a sound. For example, when an animal walks around, eats, breathes or performs any bodily function, non-intentional sounds are produced (which may or may not be useful as far as that animal is concerned, but may be decidedly useful to a second animal preying on the first). Conversely, an animal may produce intentional sounds (perhaps for the purpose of communicating with other animals) by causing parts of its body to vibrate. For example, many animals can generate sounds by laryngeal vibrations as they breathe.

Intentional sounds are usually produced for communicative purposes between animals (Richards, 1985). For instance, animals may make sounds to keep in touch with their young, or to attract mates. In humans, this communication by sounds is highly developed and is the phenomenon we know as speech. The analysis of speech sounds, which is a major concern of this thesis, is the subject of Chapters 2 through 5. Certain

animals, typified by some species of bats and aquatic mammals, employ sounds to orient themselves in their environments. These mammals are typically able to navigate and locate objects with remarkable precision merely by echo-location . An analysis of the echo-location capabilities of the sounds emitted by Hector's dolphin is presented in Chapter 7.

Non-intentional sounds contain information about the activities that produce them. When an activity is obscured in some way (e.g. because it is hidden by an opaque (to light) barrier), the sound may be the only indicator of the presence or characteristics of that activity. This ability of sounds to "reveal the hidden" is utilised by certain animals who hunt other animals by sound rather than by sight. It is also useful in medical diagnosis, because many important organs of the anatomy, although hidden inside the body, produce sounds which can be heard outside the body. Such a sound often changes its character if the bodily organ producing it has some pathological condition. Thus, by listening to sounds produced by various parts of a body, especially the heart, lungs and stomach, a trained practitioner may be able to diagnose a pathological condition of an internal organ. The lungs and airways emit sounds which change in well understood ways for pathological conditions such as pneumonia and asthma. Chapter 6 describes the sounds produced as a person coughs, and reports my investigations into the changes that occur in the sounds when the person is afflicted with asthma.

1.4.3 Speech material for signal processing

It is convenient to introduce here the speech material used to test the algorithms introduced in Chapters 3, 4 and 5. Several different utterances were employed (an utterance is a particular phrase spoken by a particular person in a specified manner). The utterances are listed in Table 1.2. Each utterance is identified by a code so that it can be referred to when required. The code identifies the person (first two letters), the person's sex (third letter), the phrase, and the manner in which it was spoken (last letter). Thus the code "AM-RAIN1" identifies an utterance of the "Rain" phrase (see Table 1.3) spoken by the (male) speaker A. The suffix "1" indicates that this is the first example of several similar utterances.

1.4.4 Procedure for recording speech sounds

Some of the utterances described in §1.4.3 were recorded in an anechoic chamber, while others were recorded in a large laboratory, in which (quiet) background noise was present. Table 1.2 indicates the particular recording location for each utterance. All utterances were recorded onto chrome-dioxide tape by an AIWA F990 cassette-tape recorder. Dolby-C was used to reduce the recording noise and an AIWA CM-53 microphone was employed as transducer. All recordings were later digitised at a sampling rate of 10kHz by a 12 bit LPA11-k A/D converter. A KEMO VBF/8 low-pass filter, having a 3dB cutoff frequency of 4.5kHz and an attenuation that increased at 48dB/octave above this frequency, was used to avoid aliasing the signals as they were digitised.

The frequency response of the microphone is flat (to within 3dB) between 50Hz and 13kHz. Likewise, the tape recorder has a flat transfer function between 20Hz and 18kHz. The phase response of the tape recorder was found to be linear, to within 10°, between 100Hz and 5kHz. The phase response of the recording apparatus is important for some of the processing techniques described in later chapters. This is because the shape of a waveform depends critically on the phases of its spectral components. §1.4.4.1 describes the technique employed to determine the phase response of the recording apparatus.

Speaker ID	Sex	Phrase	Manner or example number	Duration (s)	Anechoic Chamber?
A	M	RAIN	1,2	4.6	Y
A	M	WAL	1	8.3	Y
B	M	WAL	1	7.1	Y
C	F	WAL	1	7.6	Y
K	F	WAL	1	6.4	Y
T	F	RAIN	1	5.3	Y
T	F	WAL	1	7	Y
W	M	BRIT	1	14	Y
W	M	WAL	1	7.8	Y
W	M	RAIN	1,2,3	5	Y
W	M	RAIN	T - Tense	5.3	Y
W	M	RAIN	R - Relaxed	5.8	Y
W	M	VOWEL	(phoneme-specific)	2	N
W	M	NASAL	(phoneme-specific)	1.7	N
W	M	TESTA		2.8	N
W	M	TESTB		3.5	N

Table 1.2. List of utterances employed in the speech processing sections of this thesis. The phrases are identified in Table 1.3.

Label	Phrase	Source
RAIN	When sunlight strikes raindrops in the air, they act like a prism, and form a rainbow.	Fairbanks, 1940
WAL	'The time has come', the walrus said, 'to talk of many things: Of shoes, and ships, and sealing wax, of cabbages, and kings.'	Carroll, 1898
VOWEL	Why were you away a year, Roy.	Huggins and Nickerson (1985)
NASAL	Nanny may know my meaning.	
TESTA	The little blankets lay around on the floor.	
TESTB	The trouble with swimming is that you can drown.	
BRIT	Stitching London together, from one bank of the Thames to the other, are thirty two bridges. Twenty for road traffic, ten for rail, and two for pedestrians only. Without these bridges, Britains capital could not function.	

Table 1.3. The phrases corresponding to the utterances listed in Table 1.2.

1.4.4.1 Phase response of recording apparatus

The phase response of the recording apparatus was estimated with the use of a HEWLETT-PACKARD 3561A Dynamic Signal Analyser. This machine invokes FFT processing to perform spectral analysis of signals. In addition, pre-defined signals can be generated so that the transfer function of a system can be determined (in the manner described in §1.3.2).

In order to determine the frequency response of the tape recorder, pulses having a flat spectral magnitude across the analysis frequency range are generated by the signal analyser. These pulses are recorded on a Chromium Dioxide tape, using the Dolby C method of noise reduction. In addition, they are simultaneously analysed by the signal analyser in order to determine their spectral content. The recorded pulses are subsequently replayed and their spectral content is also computed. The phase response of the tape recording and play-back process is obtained by subtracting the spectral phase of the original pulses from that of the recorded pulses. The magnitude of the transfer function is likewise obtained by dividing the spectral magnitudes of the two signals. Note that a Wiener constant is not necessary because the spectral magnitude of each of the original pulses is almost flat across the entire range of frequencies that are analysed.

To reduce the effects of noise on the spectral estimates, fifty pulses were synchronously averaged. Each pulse was aligned by means of a trigger, set to 0.25 times the maximum input amplitude. The analyser sets a window about each pulse such that the instant at which the signal first exceeds the trigger is centred. This means that the fifty pulses constructively reinforce, while noise, such as any mains "hum", is cancelled out. Note that the duration of the analysis window varied according to the required analysis bandwidth. The actual duration for each case is mentioned in the results presented in the next paragraph. The average pulse is multiplied by a Hamming window before its spectrum is computed.

Fig.1.10 shows the phase and magnitude components respectively of the tape recorder's transfer function when computed over a range of 0–5kHz. In this case the length of the analysis window was 120ms. The phase curve has a linear slope because of the unavoidable slight difference in the triggering instant for the original and recorded pulses. However, as indicated by the straight line drawn on the curve, the phase is linear to within 10°over the frequency range 0–5kHz. In order to examine the phase response more closely at low frequencies, the analysis was repeated with a window length of 600ms, resulting in a frequency resolution of 3.75Hz between 0–1kHz. The associated phase response is shown in Fig.1.11, which indicates that the phase is linear, within 5°, between 200 and 1000Hz. At 100Hz, the phase is only 20° different from that at 400Hz.

The phase and magnitude components of the transfer function over the frequency range 0–25kHz are shown in Fig.1.12 (in this case the analysis window length was 24ms). The magnitude of the transfer function is flat to within 6dB (which is equivalent to 3dB in terms of signal power) up to a frequency of 18kHz. However, the phase is significantly non-linear over the entire range. Because the only acoustic energy utilised was that for frequencies less than 5kHz, the non-linear phase for higher frequencies is of no importance.

The accuracy with which the recorder preserves the shape of the waveform for a typical segment of voiced speech is illustrated in Fig.1.13. A speech utterance was recorded on the tape and simultaneously digitised directly onto the computer. The recorded version was subsequently replayed and also digitised. Figs.1.13*a* and 1.13*b* show the directly digitised and recorded versions respectively. The average power in the error between these two signals is of the order of 30dB less than the average power

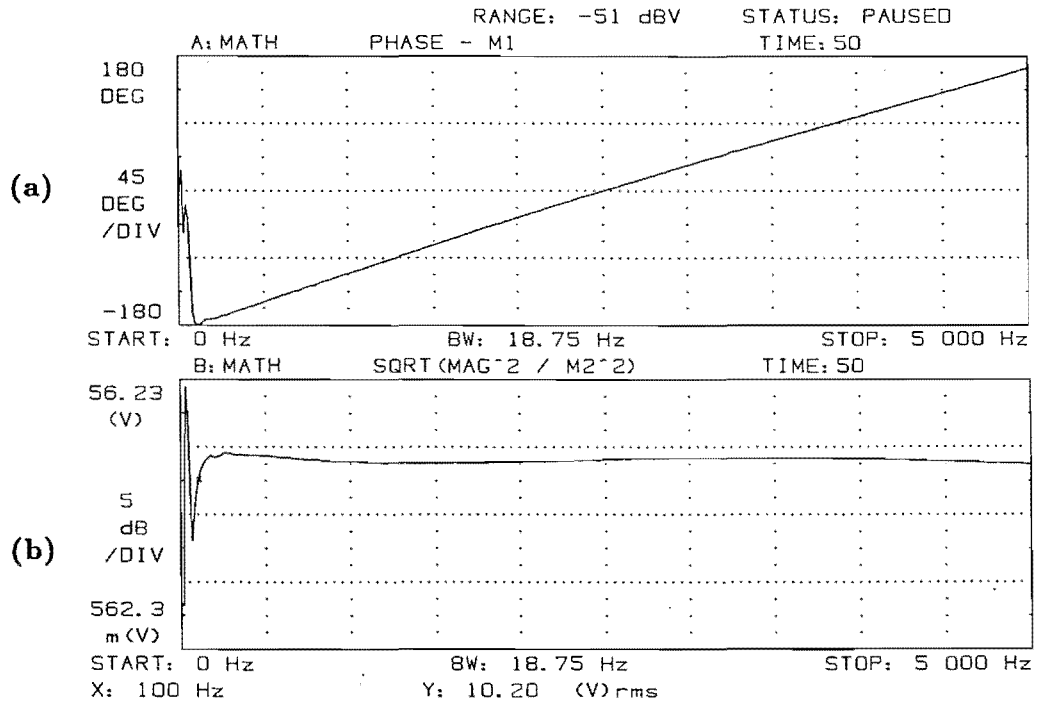


Figure 1.10. Transfer function of tape recorder, computed over a frequency range of 0–5kHz by a HEWLETT-PACKARD 3561A Dynamic Signal Analyser. **a:** Phase, and **b:** magnitude, of transfer function.

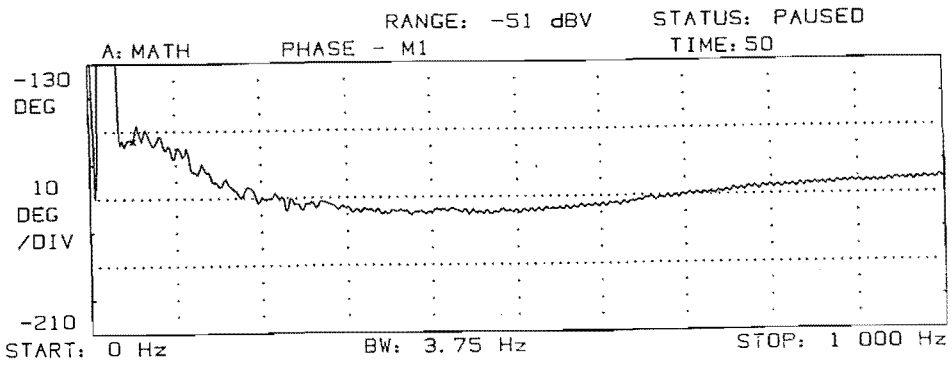


Figure 1.11. Phase of the transfer function of tape recorder, computed over a frequency range of 0–1kHz.

in the speech signal itself.

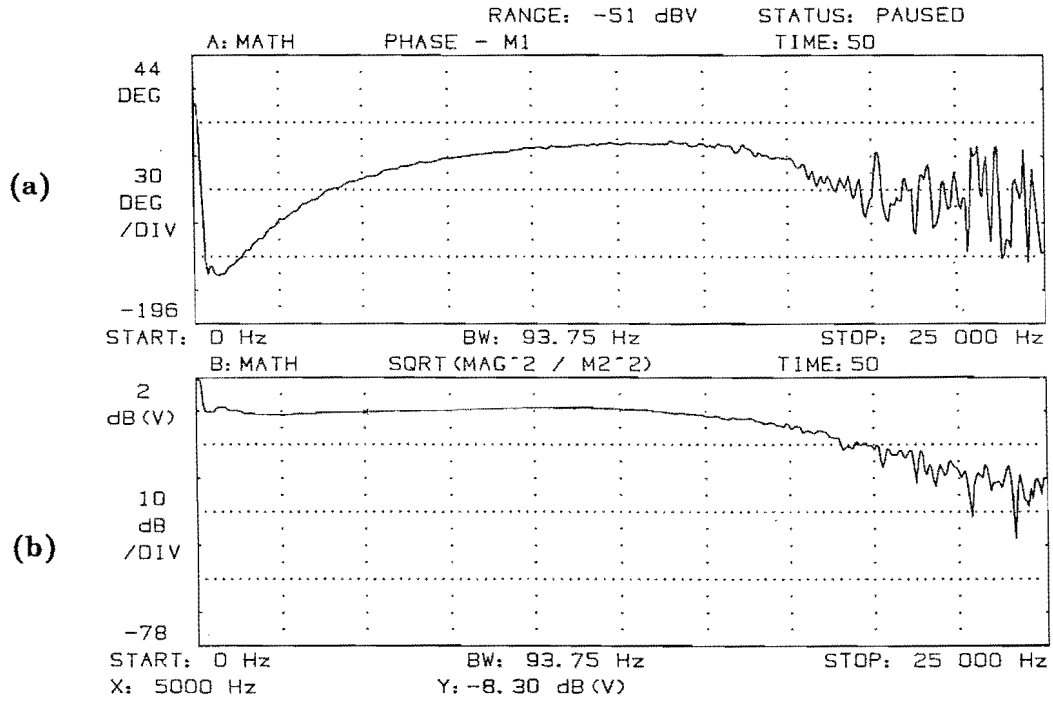


Figure 1.12. Transfer function of tape recorder, computed over a frequency range of 0–25kHz. **a:** Phase, and **b:** magnitude, of transfer function.

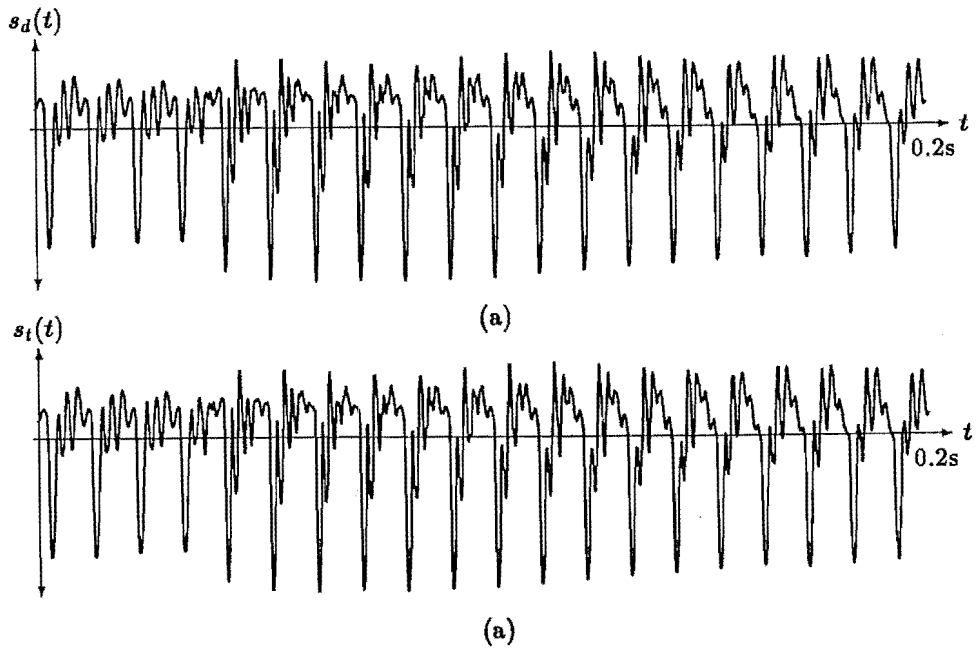


Figure 1.13. Illustration of effect of tape recorder on the waveform of a typical speech signal. **a:** Directly digitised version of speech signal. **b:** Recorded version of the same speech signal as in *a*. The signal represents part of the word "Hello", and is digitised at a sampling rate of 10kHz.

Chapter 2

Speech sounds

This chapter introduces speech sounds, from the points of view of their use in human communication, the way in which they are produced, and how they are modelled to facilitate analysis. After touching on the uses that speech has as a means of human communication (§2.1.1), I discuss some of the motivations for undertaking research into speech analysis techniques. In §2.1.3 through §2.1.4 I outline the basis of the linguistic and phonological approach to analysing speech sounds, and discuss the implications for technological speech analysis that arise from such a study. Section 2.2 contains a review of the human speech production and perception mechanism, while §2.3 is an introduction to some models of speech sounds which provide the bases for the analysis techniques discussed in subsequent chapters.

2.1 Introduction to speech sounds

2.1.1 Speech and communication

The primary purpose of speech is to serve as a means of communication between people. Therefore, as an introduction to speech signals, I briefly discuss the issues raised by the term communication and how they apply to speech.

Communication involves the exchange of information between individuals who are able to process that information. These individuals, who in general might include species of automatic systems, must be connected by some channel through which the information can travel. In almost all cases, the information must be transformed from its internal representation, by means of which each individual or system stores it, into a form suitable for transmission along the channel. For meaningful communication to occur, the individuals involved in the communication must agree on protocols for representing the information in the channel and for how it is to be transformed between their internal representations and the channel representation.

In the context of communication between people, the information to be exchanged consists of thoughts, ideas, emotions, instructions, etc (Cherry, 1978). The communication channel is a medium through which can travel some signal corresponding to one of our senses (since it is our senses which enable us to receive information from outside our bodies). For most communication purposes, channels consist essentially of sounds (to accord with our sense of hearing) or images (corresponding to our sense of sight). Note that although, strictly speaking, the channel is actually the air or space (and any intervening technological apparatus) through which the signal (sound or image) travels, it is convenient here to consider the sounds or images to be the channel. The mechanisms by which sounds actually travel between people is described in Chapter 1. Information is represented in sound form by means of speech, music,

or other codes (such as Morse code), and in image form by means of pictures, body-language, symbols, or written representations of sounds (such as writing). Obviously, the people who are attempting to communicate must "speak the same language" before their efforts can be successful.

Speech is one of the most important methods of communication between humans. In addition it is employed by all known human communities as a basic form of communication. Its importance arises from several factors. Firstly, speech communication is applicable to a wide variety of messages, ranging from subtle emotions to precise instructions. Secondly, the information in speech is encoded in terms of a well defined language. Language is a term which is often taken to be synonomous with speech, but in a more general sense, it can be defined as any means of expressing or understanding thoughts, ideas or emotions (Skinner and Shelton, 1978, p 8). Correspondingly, other communication forms (such as art or music) can be said to be languages, although usually of more amorphous kinds. Spoken language has widely accepted rules which specify its structure and how it is to be used and interpreted, so that the scope for errors (misunderstandings) occurring during the communication process can be minimised. Of course the rules vary enough from place to place and time to time that misunderstandings do occur easily enough, as Molinger (1975, Chapters 11 to 13) emphasises when discussing the dynamics of language variation. The role that language has in encoding the information in speech is discussed more fully in §2.1.3.

Another aspect of language is that it has a profound effect on how we think, so that we organise our ideas to conform to the structure of language, and our thoughts often proceed in terms of an *internal speech.* There is still debate about the relationships between thought, language and speech (Cutting and Kavanagh, 1975; Molinger, 1975, Chapter 8), but whatever their relationships, it is enough to say here that they are intimately entwined. This means that the information which we may wish to communicate by some other method (such as by Morse code or smoke symbols) is very often already encoded in terms of speech, and hence these other methods are often just ways of encoding speech in another medium.

A further indication of the importance of speech as a means of communication is that it can be directly transcribed into its symbolic representation, writing, by means of which speech sounds can be permanently recorded. Any person who knows the transformation (i.e. who can read) can then reconstruct the original words, and hence the ideas that they represent. Of course, some of the information inherent in spoken language, such as the speaker's voice characteristics, or how the words were spoken, cannot be represented in the written form of the language, except by means of extra instructions, and even then only approximately (such as *"Look at this one!", he gasped excitedly.*).

2.1.2 Technological extensions to speech communications and motivations for speech analysis

As emphasised in the previous section, the role of speech as a means of communication is important to humans. By means of appropriate technology, its applicability can be extended beyond its most basic form of two individual (or groups of) people talking to each other. Various forms of technology can enable us to use speech to communicate over long distances, or to many people at once, or to help us remember things for long periods of time. We could even communicate with machines using speech rather than by turning knobs and pressing buttons.

The first extension is that mentioned previously of writing down a symbolic representation of the speech sounds (using simple technology such as paper and pencil). This extension is extremely well developed (in the major languages), so much so that

in many cases the speech sounds are never actually spoken, but are transcribed directly from the thoughts of the writer (the thoughts generally being in terms of a language) into written words, and then directly into thoughts by the person who reads the words. Writing can be stored for a very long time or transported to someone at a distant location (who must of course still know how to read the same language). It can also be duplicated many times without loss of information (by means of a technology such as printing) and thus the information can be widely disseminated.

More recent technological advances (such as radio, telephone and sound recording) have enabled the speech sounds themselves to be stored, duplicated and transmitted great distances, without being first transformed into their written form. The popularity of the devices performing these tasks is another testimony to the importance of speech as a means of communication between people. In fact, the demands on both speech storage systems and telephone networks is such that new technologies are constantly being devised in order to improve their performance. In the case of speech storage systems, people want to store more speech, and gain access to any part of it more quickly, than the "old" technology of tape-recorders is capable of. In the case of telephone networks, people make more and more telephone calls each succeeding year, and wish to do so with greater flexibility (using portable phones for instance). To realise the desired improvements, techniques are needed to compress speech signals for storage or transmission, in such a way that close replicas of the original sounds can be reconstructed later. This is the basic motivation for developing the speech coding algorithms which are discussed in Chapters 3,4 and 5 of this thesis (cf. Atal and Rabiner, 1986;Andrews, 1984).

The development of computers has led people to enquire whether they can be made to "understand" human speech, and/or "speak to us". The two main areas where this may be (and, in fact, already is to some degree) useful are, firstly, controlling and communicating with machines (including computers), and secondly, transcribing speech sounds into their written equivalent and vice versa (Fant, 1985; Atal and Rabiner, 1986; Immendörfer, 1986; Andrews, 1984).

Speech recognition is currently a very active field of research (Atal and Rabiner, 1986; Jakatdar and Mulla, 1986) and although the actual recognition algorithms are outside the scope of this thesis, most of the speech analysis techniques that I discuss in later chapters can be invoked during the feature extraction stages of the recognition process. Likewise, in the text-to-speech or machine output process, the initial translation from text to a representation of the speech sounds is in the realm of expert systems (cf. Fant, 1985). However, the synthesis of the speech sounds is based on the same techniques which I discuss in succeeding chapters.

The production and perception of speech by humans are processes which are still not fully understood. One of the ways in which our knowledge of these processes can be increased is by means of signal processing techniques (cf. Boves, 1984). The speech production process can be modelled (§2.3) and the speech signal analysed in such a way that the contributions of various parts of the speech production mechanism can be isolated and studied in detail (Fant, 1973). Furthermore, the effect that various pathological conditions have on the speech signal can be identified. This use of speech analysis for diagnostic purposes is another area wherein there is currently much research (Kasuya *et al.*, 1986; Childers *et al.*, 1986).

2.1.3 Information carried by speech sounds

The information in the speech signal consists of the linguistic message, together with *para-linguistic* information such as whether the message is a question or statement, and *non-linguistic* information, which includes the speaker's emotional state and identity

(Laver, 1980). In any analysis of the speech signal, the information can be considered as being encoded at various levels, ranging from the physical structure of the acoustic signal to the actual ideas and emotions imparted by the signal.

In §2.1.3.1 through §2.1.3.3 I examine the various intermediate levels at which the information can be considered as being encoded. The discussion is limited to those aspects having an impact on speech analysis systems. The linguistic and psychological implications, which are outside the scope of this thesis, are very extensive and have been elucidated in depth by many people (cf. Wickelgren, 1976; Molinger, 1975; Jakobson, 1978; Lieberman and Blumstein, 1988; Skinner and Shelton, 1978). For each of the "levels" described, I discuss how the information is encoded, and how the level relates to the "higher" and "lower" levels.

2.1.3.1 The acoustic, or phonetic level

At the physical level, speech consists of a time-varying acoustic signal, with a spectral content ranging from about 50Hz to 10kHz. The information is carried by the manner in which the spectral content of the signal changes with time. However, the structure of the signal is very complicated, and it is useful to describe it in terms of various features, called phonetic features because they relate to the actual sound. The information can then be described by the way that the features change, although a problem arises in the choice of features, since some features of the sound carry little linguistic information, but may appear to alter the signal markedly. The choice of features to describe the signal must therefore be guided by knowledge of their linguistic significance, as is emphasised by Jakobson's (1978) Lecture 1 and Fant's (1973) Chapter 1 .

The relative importance of various phonetic features for describing speech sounds has been established by many investigations (cf. de Saussure, 1959; Jakobson *et al.*, 1961; Fant, 1973; Boves, 1984). The most important features for linguistic purposes are the spectral content of the signal (§2.3.1.3), type of excitation (§2.3.1.1) and the timing between various events. Other features such as the pitch frequency (§2.3.1.1) and loudness are of less importance for the linguistic message (at least in English).

Non-linguistic information includes the emotional states of speakers, and the various types of information that serve to identify individual speakers (Laver, 1980). For example, the pitch frequency (see §2.2.1) can indicate the sex or age of the speaker. The emotional state of the speaker is largely indicated by changes in the prosodic structure of their speech (Williams and Stavens, 1972). Other types of non-linguistic information include the accents of speakers, which indicate their cultural origins, and may also indicate the social context in which they are talking. Non-linguistic information is encoded in the long-term average of the phonetic features, and in the specific "peculiarities" of the ways that the phonetic features are varied (see §2.1.4.1; Boves, 1984; Laver, 1980).

Together with the loudness of a sound, the pitch frequency can give an indication of the emotional content of the message being communicated, as well as carrying information about the type of message (e.g. whether it is a question or an insult). This is termed para-linguistic information, and also includes information about the overall structure of the linguistic message, such as where sentences begin and end (Lieberman and Blumstein, 1988, pp198–203). Para-linguistic information is essential as an aid to understanding the linguistic information implied by speech sounds. Its usefulness is clearly demonstrated in situations where we can hear the sounds of someone talking in the distance, and although we cannot understand the words, we can often guess what they might be talking about.

Linguistic information is encoded in the way that the phonetic features, especially the spectral content and excitation type, change with time (Peterson, 1952; see also §2.1.4.2). However, there is not a one-to-one relationship between the features and the linguistic meaning. A particular set of phonetic features may have different linguistic meanings depending on the surrounding context (such as which words precede or follow it, or even the social context in which they are spoken). Furthermore, there is a great deal of variability in the exact way that words are pronounced, both by different speakers, and by the same speaker at different times (Fant, 1973, p19).

2.1.3.2 The phonemic and linguistic levels

To overcome the problems (outlined in the previous section) of associating acoustic and linguistic features, the concept of a phoneme has been developed (see the most illuminating discussion by Jakobson, 1978). A phoneme is defined as a member of a set of sound units from which all words and sentences in a particular language can be constructed, by concatenating various phonemes. Each phoneme is associated with various acoustic features but, as implied in the previous section, these vary according to the specific context in which they occur.

Phonemes may be said to be the blocks from which words and sentences are built. However, each language has a unique set of phonemes (Jakobson, 1978, Lecture 2). This is because phonemes are defined from a linguistic view-point, in that different phonemes are different only in that their use alters the meaning of a word in that language. Hence, some particular sounds may be heard as clearly different phonemes to a speaker of one language, yet they may merely be variations of the same phoneme, in some cases indistinguishable from each other, to a speaker of another language. The different acoustic forms of a particular phoneme, which may occur in different contexts, are termed *allophones*.

The implication of the above definition of a phoneme is that the simplistic association of a phoneme with a particular set of acoustic features is inadequate. The particular acoustic features that distinguish a phoneme from other phonemes in a language may change dramatically according to factors such as the phoneme's position in a word or the social context in which the speaker is talking. This is further exacerbated in more informal contexts where people may abbreviate or even skip certain phonemes in a word. The important point is that every phoneme has some different acoustic features in each particular context (Jakobson *et al.*, 1961, p4). Human listeners can use their knowledge of contexts to help them decode speech sounds into phonemes, and thence into words.

Notwithstanding the above statements about the variability between different occurrences of a phoneme, some phonemes, especially the vowels (§2.1.4.2), have well defined acoustic features which are reasonably invariant (for a particular speaker) in different linguistic contexts (cf. Fant, 1973; and chapters 8 and 10 of Lieberman and Blumstein, 1988). These features relate to the spectral shape of the sound, and especially to the positions and amplitudes of the various *formants*, or spectral peaks (see §2.3.1.3). The consonant phonemes are more variable, in that while their acoustic manifestations vary according to which phonemes precede and follow them, a human listener perceives the same "sound" for each combination. A description of the various types of phonemes and their acoustic representation is presented later, in §2.1.4.

Any particular phoneme has no meaning in itself, other than to be different from other phonemes and hence to act as a distinguishing feature between different words (Jakobson, 1978, Lecture 4). That the phonemes have no intrinsic meaning is demonstrated by considering pairs of words that differ only by one phoneme, such as *hat*, *hot* or *cat*, *cot*, but differ widely in their linguistic meaning. When the speech

signal is considered as a sequence of phonemes, the information is encoded in how the sequence is ordered. Particular sequences of phonemes correspond to words and sentences. This is termed the linguistic level, and the information can be understood as a series of words. Like the transformation from sound to phoneme, a particular sequence of phonemes may not uniquely determine a particular linguistic meaning. However, the linguistic content is almost always sufficient to resolve the ambiguities.

Words themselves only contain information because they are symbols which signify some idea or object (Wickelgren, 1976). This is the level at which the meaning, implicit in the lower levels, is made explicit. However, the transformation between a word and the conceptual idea that the word signifies is again subject to ambiguities, which can usually be resolved by recourse to the surrounding context (and which in certain instances can give rise to humorous puns).

2.1.3.3 Implications for speech analysis schemes

The discussion in §2.1.3 of how information is encoded emphasises that the information content of a speech signal is only really unambiguously understood when it is decoded at the conceptual level. Only at this level is the entire context in which the communication is occurring (consisting of the particular language, the linguistic context and the social context) known.

The problems involved with describing phonemes in terms of unique acoustic features impose the major constraints on any system of speech analysis where the aim is to extract the phonemes and/or words from the sounds. Without access to higher-level information concerning the linguistic and social contexts, such a system cannot successfully match the acoustic sounds to a unique set of phonemes or words.

When compressing speech signals for storage or transmission, we are concerned with retaining the information in the signal which pertains to the particular application, but we are happy to dispose of any extraneous information. The linguistic information must obviously be retained, and this involves preserving those features of the speech sound which distinguish between phonemes. Particular attention must be paid to those phonemes in a language which differ only slightly in their acoustic manifestations. It is also relevant to note here that, since different languages tend to have different sets of phonemes, a speech compression system designed for one language may perform better or worse when it is applied to speech of another language.

For compression systems which are designed to produce natural sounding speech, the acoustic features which carry the non-linguistic and para-linguistic information must also be preserved. This includes those features which identify the voice characteristics of the speaker, and features such as pitch and loudness which also convey some information about the message. The problem is to preserve the "natural soundingness" of the speech, while still achieving whatever level of compression is desired (Laver, 1980).

The success of any scheme designed to generate intelligible and/or natural sounding speech at a particular level of compression can only be evaluated subjectively by human listeners. Thus it is useless to assess a speech compression scheme by comparing the input and output signals on a simple acoustic level (§3.5.3). The information in the signal retained by the system can only be assessed by the response of human listeners to the output sounds. This is effectively a re-iteration of the assertion made earlier in this section, that the information in a speech signal can only be understood in its full context.

2.1.4 Descriptions of various types of speech sounds

Having introduced the concept of a sound unit, or phoneme, in §2.1.3.2, it is now useful to briefly discuss the different types of phonemes, both the ways in which they are produced, and the various acoustic features that identify them.

2.1.4.1 Speech quality

One factor which is easily overlooked in a description of different speech sounds is that phonemes can only be described as relative to each other in a particular context. For example, each person has a different "voice quality", and so the actual acoustic representation of a phoneme differs according to the person who utters it (Laver, 1980). However, the relative differences between phonemes, whether measured in articulatory or acoustic terms, are likely to be similar for different people (Fant, 1973). In addition, people who speak with different accents may have quite different ways of producing a certain sound, but human listeners are still able to identify the similarities and differences between phonemes uttered by such different people. For instance, Fig.2.1 shows the vowels of both New Zealand and North American speakers graphed with reference to their first two formant frequencies. Despite the major differences in the positions of the formants of each vowel (cf. Hawkins, 1973; Maclagan, 1982), North Americans and New Zealanders are still able to understand each other most of the time. This ability is due to the contextual redundancy inherent in speech (§2.1.3), whereby a person can determine the identity of phonemes and words even though they are "mis-pronounced".

The types of features which give voices distinctive qualities arise from peculiarities in types of laryngeal vibration (cf. Laver, 1980, Chapter 3; Boves, 1984, Chapter 5) or because of variations in the ways in which people configure their vocal tracts to produce particular sounds (Laver, 1980, Chapter 2). The long term characteristics of a person's voice are called "articulatory settings" (Laver, 1980, p12) which is a term used to describe the "average" sound, or 'the common, rather than the distinguishing components' of sequences of phonemes (Laver, 1980, p13; Lieberman and Blumstein, 1988, pp152–154). Styles of pronunciation are also characterised by peculiarities in the prosodic structure of people's speech (e.g. pitch or tempo variations).

Peculiarities in laryngeal vibration are also affected by any pathological dysfunction of the larynx, and hence this aspect of speech quality, when it can be usefully described and quantified, can also be useful for diagnostic purposes (see §3.6.4.1).

2.1.4.2 Phonetic and Phonemic classification

Different phonemes can be classified according to the ways in which they are articulated, their acoustic manifestations, or how they are perceived by human listeners (Edwards and Shriberg, 1983, p12). The classifications are necessarily based on the differences between phonemes, since as noted in §2.1.4.1 phonemes uttered by different people are different in an "absolute" sense, but "relative" to each other are similar for each person (Lieberman and Blumstein, 1988, p179). As stressed in §2.1.3.2, the features must be considered in the context of an overall linguistic framework. I describe the various phonemes in terms of how they are pronounced by New Zealanders, which is different than that of other English-speaking countries (Hawkins, 1973,1976)

A set of phonemes for a particular language contains the basic "building blocks" of the language. However, in order to study the varieties of pronunciation between different groups of people, a more detailed description of the *phonetic* differences between sounds is required. The *phonetic alphabet* is a collection of symbols, each of which identifies a particular speech sound. Hence a *phonetic transcription* of a speech utterance provides a detailed description of how it was pronounced (Cherry, 1978, p80).

In any particular language, several phonetically different sounds may correspond to a single phoneme (§2.1.3.2). Each phoneme in a particular language is denoted by an appropriate phonetic symbol. In order to distinguish phonemic signs from phonetic signs, phonemic signs are enclosed by slanting lines (e.g. /n/), while phonetic signs are enclosed by square brackets (e.g. [n]) (Ladefogad, 1982, Chapter 2). A *phonemic transcription* of an utterance is necessarily less detailed than a phonetic transcription, as it retains only the bare linguistic structure of the utterance (Ladefogad, 1982, pp23–24), without the details of an individual's pronunciation of it.

A description of sounds in terms of their articulation is useful for speech-language therapists who wish to help people overcome language disabilities. However, linguists who attempt to describe the phonetics of a language are often more interested in how sounds are distinguished, and so adopt an auditory approach. The acoustic classification of speech sounds involves analysing sounds by means of technological apparatus, so as to be able to extract features characteristic of the sounds. An acoustic description is only really meaningful as an aid to one of the other two classification schemes, because of the difficulty in finding separate acoustic segments that directly relate to phonemic elements (see the discussion in §2.1.3.1, and also page 162 of Lieberman and Blumstein, 1988). I follow the articulatory classification here, since this is the more traditional approach (Lieberman and Blumstein, 1988, p163). In addition, I give a brief description of the acoustic features that typify each type of sound. Further details of the acoustic characteristics of each type of articulatory feature are presented by Fant (1968, pp236–252).

In the articulatory classification scheme, the major distinctions between different sounds are the amount of obstruction in the vocal tract and the type of excitation used to produce the sound. Generally, the primary division is between vowels and consonants, because vowels are produced with a relatively open vocal tract and consonants with more obstruction (Edwards and Shriberg, 1983, p13). Each consonant is further characterised by the position where the obstruction occurs, the manner in which the obstruction is effected (e.g. complete or partial obstruction of the airflow) and whether voicing is present (Table 2.1). Vowels are traditionally classified according to tongue height, tongue position, and lip rounding (Ladefogad, 1982). From an acoustic point of view, each vowel can be uniquely represented by two formant frequencies, although there is some debate about how the second formant is best defined to effect this characterisation (cf. Bladon, 1983).

Vowels are always voiced, and are articulated with a relatively open vocal tract shape (Lieberman and Blumstein, 1988, pp171–182). Furthermore, vowels such as /i/, /ɔ/ and /a/ (or /i/, /u/ and /a/ for British and North American English — see the discussion on different accents in §2.1.4.1) are each very different, both in the vocal tract shape which produces them, and in their particular formant frequencies (Stevens, 1989; Perkell and Cohen, 1989). If the vowels are represented by a graph of the first formant frequency versus the second formant frequency (Fig.2.1) then the above vowels are found to be at the corners of the resulting "vowel space". Even though different speakers utter these phonemes with different (absolute) formant frequencies, the (relative) differences between each phoneme are similar (Lieberman and Blumstein, 1988, p178). It seems that, when listening to a different speaker, a person uses these phonemes as reference points to "tune in" to the particular way in which the speaker talks, because of the distinctive nature of these phonemes. All the other phonemes then fit into the framework in a similar place for each speaker (Lieberman and Blumstein, 1988, pp176–182). The phonetic characteristics of different vowels are the frequencies of the formants, especially the first two. The frequencies involved are between about 200Hz and 3500Hz.

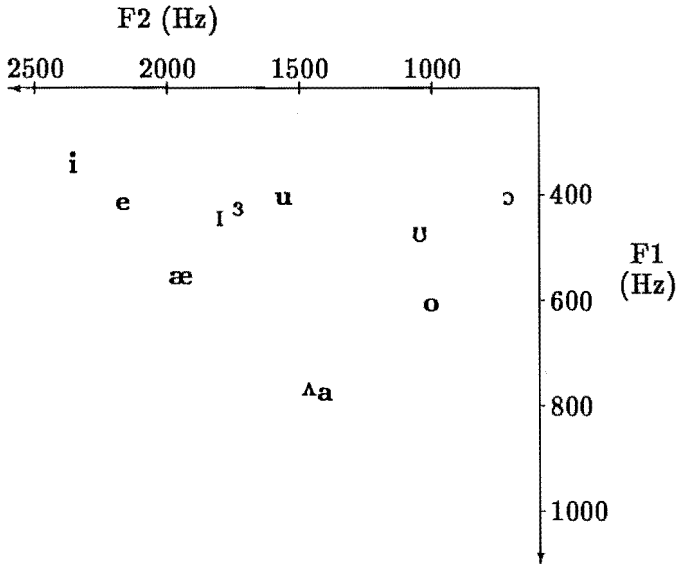
Manner of articulation	V/UV	Place of articulation (obstruction)						
		Bilabial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
stop	V	**b** bib			**d** did		**g** gig	
	UV	**p** pet			**t** tot		**k** key	
fricative	V		**v** very	ð then	**z** zoo	ʒ azure		
	UV		**f** fat	θ thin	**s** sat	ʃ sham		**h** had
affricate	V					ʤ june		
	UV					ʧ chin		
nasal	V	**m** map			**n** nip		ŋ sing	
liquid	V				**l** lull	**r** run		
glide	V	**w** why				**j** yet	**w** why	

Table 2.1. Diagram of the consonant sounds of English, classified according to their place and manner of articulation (after Edwards and Shriberg, 1983).

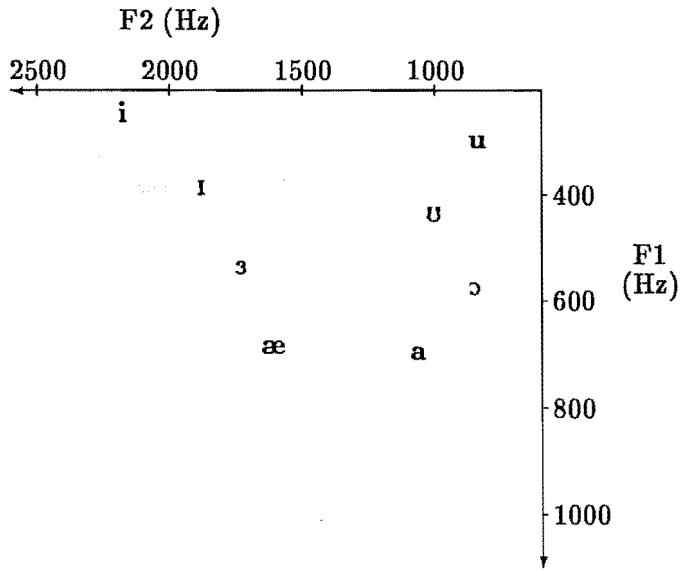
Dipthongs are vowels in which the articulatory configuration is not constant, but varies throughout the one syllable. Dipthongs are represented by the two phonetic symbols that are closest to the start and end positions of the dipthong. For example, in the word "my", the /**ai**/ dipthong exhibits a smooth transition between the /**a**/ and /**i**/ sounds (Ladefogad, 1982, pp76–78).

Consonants are more diverse than vowels, in that they may be voiced or unvoiced, and may be caused by different types of obstruction (Table 2.1). In addition, they are extremely variable in both their acoustic manifestations and their forms of articulation, depending upon the surrounding sounds. This has traditionally been termed "co-articulation", where the sounds of adjacent phonemes overlap (cf. Lieberman and Blumstein, 1988, p145).

The "stop" consonants are produced by completely obstructing the vocal tract momentarily, with the position of the obstruction altering the characteristic of the sound produced upon re-opening the tract. Stops may be voiced or unvoiced, the difference being whether the larynx begins vibrating as soon as the obstruction opens, or after. Each of these sounds exhibits a short silence period, followed by a short burst of sound. The identity of voiced stops is determined by the "transitions" that the formants describe from the onset of voicing to their steady-state values (cf. Chistovich, 1968; Harrington, 1988). This is illustrated in Fig.2.2, which shows the formant transitions that characterise the phonemes /**g**/, /**d**/, and /**b**/. The second formant drops from its initial frequency in the phoneme /**g**/, rises slightly for the phoneme /**d**/, and rises somewhat more for the phoneme /**b**/ (Lieberman and Blumstein, 1988, p144). Unvoiced stops are similarly identified, although the later onset of voicing means that the frequency content of the burst of unvoiced sound is also significant (Stevens, 1985;



a: New Zealand English



b: American English

Figure 2.1. Plot of the first versus the second formant frequencies for the vowels of **a:** New Zealand English (after Maclagan, 1982), and **b:** North American English (after Ladefogad, 1982, p179).

Harrington, 1988).

"Fricative" consonants are produced by severely constricting the vocal tract so that air turbulence is produced. The way in which the constriction is formed influences the sound of the turbulence. Acoustically, the fricatives produce a wide bandwidth "noise-like" signal which, due to the effects of co-articulation, is affected by the formants of adjacent vowels (e.g. note the difference in the /s/ phoneme in the two syllables /su/ and /si/). Fricatives can also be produced with simultaneous voicing, in which case the "noisy" sound is superimposed on the voiced sound (e.g. /z/ sounds). The nature of a fricative sound is determined by the distribution of energy in the higher frequency

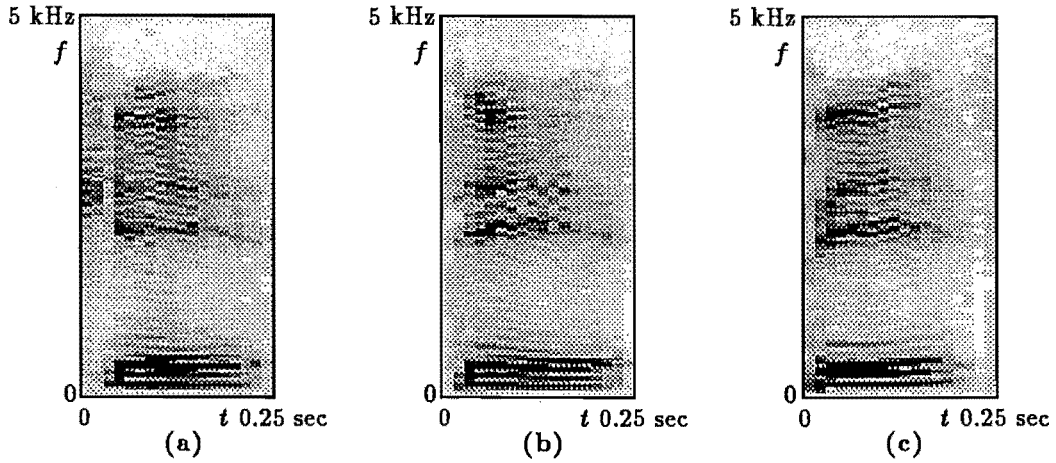


Figure 2.2. Spectrograms showing the formant transitions for the phonemes /g/, /d/, and /b/. The phonemes were spoken in the words **a:** "get", **b:** "debt", and **c:** "bet".

parts of the spectrum (Behrens and Blumstein, 1988). Voiced fricatives exhibit an additional lower frequency voiced component. The energy in fricative sounds can extend to frequencies well above 5kHz (Hughes and Halle, 1956).

"Nasal" consonants are produced by blocking the airway through the mouth and opening the velum so that the sound travels through the nasal passages. The mouth acts as a closed resonator and so affects the sound quality according to where the blockage occurs. Note that other sounds can also be produced with an open velum. This occurs especially with vowels that occur immediately before or after a nasal consonant, and is also a characteristic of some types of accents. Nasal sounds are characterised by nulls in their spectra, caused by acoustic energy lost in the closed resonator formed by the shut mouth (cf. Stevens, 1985).

The final class of consonants is the "glide" or semi-vowel, for which the vocal tract is not open enough for the sound to be termed a vowel. However, for purposes of acoustic analysis, semi-vowels can be treated in the same way as vowels, with the formant frequencies determining their identity.

2.1.4.3 Implications for speech analysis schemes

All the acoustic features of speech sounds are contained in the band of frequencies below 10kHz. Furthermore, most features, apart from some of the identifying features of stops and fricatives, are distinguished by frequency components below 4kHz. Thus, provided that the signal frequency components below 4kHz are faithfully maintained, virtually all the phonetic information in the speech signal is retained. In telephone systems, speech signals are band-limited to between 300 and 3400Hz with only a small amount of noticeable degradation.

Systems that analyse speech for compression or recognition purposes do not store the entire spectrum. They extract features which describe speech sounds, and so the ability of the features to adequately describe the information in the speech signal must be considered. In addition, the robustness of the feature extraction scheme is an important aspect, since speech signals exhibit stochastic variations, and they are almost always corrupted with significant amounts of noise (from other sound sources in the environment).

Analysis procedures for compressing and reconstructing speech must preserve those features that serve to distinguish between different phonemes (§2.1.3.3). These

include the frequencies and amplitudes of the first two formants (the other formants have much less perceptual importance), the type of sound excitation, and information related to the occurrence and timing of stops. If the aim is to produce natural sounding speech, retaining some of the characteristics of the speaker's voice, then features such as the peculiarities in the glottal excitation (cf. Holmes, 1973) and the higher order formants must also be preserved

For speech recognition purposes, analysis of the speech signal must be capable of separating out those components of the signal which relate to the characteristics of the speaker's voice from those that describe the linguistic information. This task is made easier for a recognition system if it is restricted to recognising words from a single speaker only. For recognising words or phonemes from different speakers, the system must be able to account for the differences in pronunciation between different speakers. For example, it might derive the "vowel space" of each particular speaker (Fig.2.1), and recognise the vowels on the basis of their relative position in that space. For recognising phonemes (especially consonants), the system must have stored within it all relevant information concerning how any phoneme's features change when associated with other phonemes.

2.2 Speech and Hearing

Speech consists of a coded sequence of sound patterns. It is produced (in humans) by the common activity of talking, and received by means of hearing.

Talking comprises the neurological processes of transforming thoughts into nervous impulses (to control the vocal apparatus) together with the physiological and mechanical processes of sound production itself. The first step in the act of talking consists of formulating a linguistic (consisting of words and sentences) message which expresses the thoughts and ideas that the speaker (the person who is talking) wishes to communicate. This is a cognitive process, and for further details, readers are referred to any textbook on psycholinguistics (cf. Carterette and Friedman, 1976). In §2.2.1 I discuss the physiology and mechanics of how sounds, and especially speech sounds, are produced.

Hearing is integral to the speech communication process, both because talking would be pointless without it and because of the feedback role it plays in the speech production process (cf. van Riper and Irwin, 1958, Chapter 6). Hearing consists of the mechanical processes of transforming the sound patterns into nerve impulses, followed by the cognitive process of decoding the linguistic message. In §2.2.2 I briefly outline the physiology and function of the ear, and discuss some of the characteristics of sound and speech perception by humans.

2.2.1 Speech production physiology

Humans produce speech by means of the so-called vocal apparatus, shown in cross-section in Fig.2.3. This consists of parts of the anatomy that have a primarily life-supporting role, but that have been adapted to facilitate speech production (Lieberman, 1975). It is convenient to partition the vocal apparatus into three anatomically-based parts: sub-laryngeal, laryngeal and supra-laryngeal (Fig.2.3).

The vocal apparatus is controlled by the brain via the nervous system. Various parts of the brain have been identified as contributing to speech production processes. However, in what follows I simply assume that the nervous system exists and is able to control the vocal apparatus in the ways described. A thorough introduction to the neurophysiology of speech production is given by Hardcastle (1976, Chapter 1).

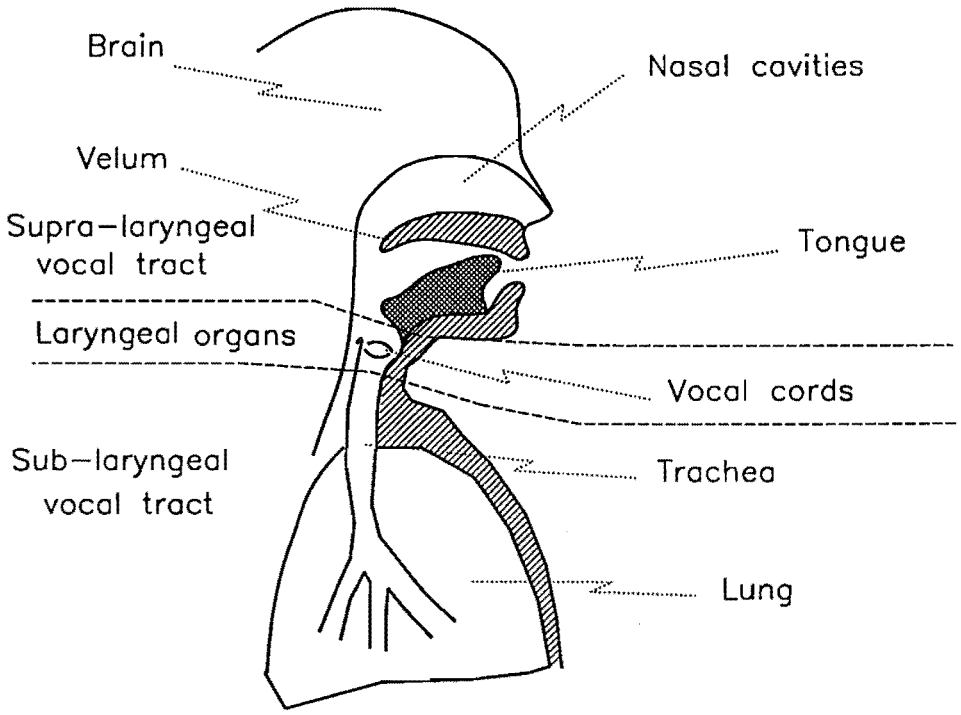


Figure 2.3. Cross-sectional view of the vocal apparatus (after Flanagan, 1972).

An important characteristic of this control system is that it is a closed loop system, with feedback coming both via the sensory nerves in the muscles themselves and from actually hearing the sound as it is uttered (Skinner and Shelton, 1978; Hardcastle, 1976, pp13–32). Impairment of any part of the feedback-control loop can have a detrimental effect on a person's ability to speak (van Riper and Irwin, 1958, Chapter 6).

2.2.1.1 The sub-laryngeal vocal apparatus

The sub-laryngeal (below the larynx) part of the vocal apparatus consists of the lungs, trachea and associated airways, and the various muscles which are used for breathing. The primary purposes of these organs are to transfer oxygen to the blood and carbon dioxide from the blood to the atmosphere. However, in doing so, they produce a flow of air through the larynx. This provides a source of energy from which sound can be generated. The natural function of breathing can be modified by humans so that the expiratory flow of air is held at a relatively constant rate for comparatively long intervals (e.g. tens of seconds) while the inspiratory phase of the breath cycle is shortened to less than a second (Lieberman and Blumstein, 1988; Hardcastle, 1976). In some people, such as singers or public speakers, this breath control is developed to a very high degree.

2.2.1.2 The larynx

The air from the lungs passes through the trachea to the larynx and thence out of the mouth and nose (Fig.2.3). The larynx consists of several cartilages surrounding the airway, together with the "vocal cords" which can be adjusted so as to obstruct the airway appropriately (see Fig.2.4) and hence generate sounds. The cartilages serve mainly to anchor and protect the various muscles which control the constriction. The

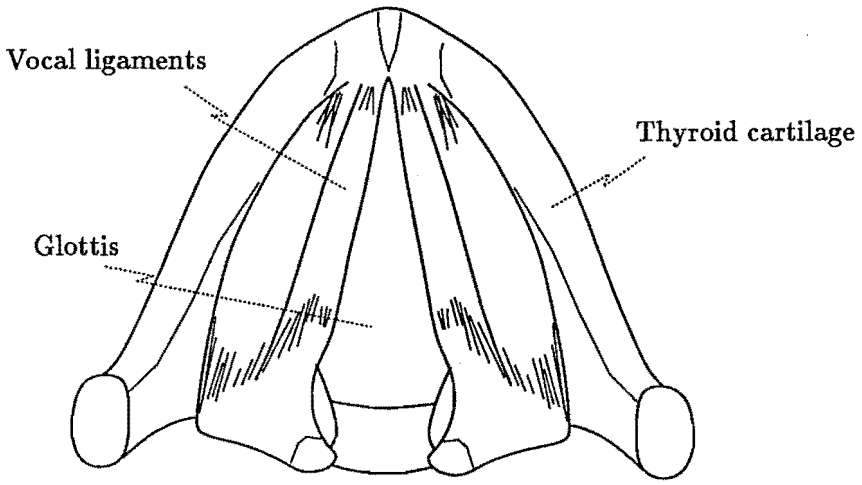


Figure 2.4. Diagramatic view of the larynx (after Gobl, 1989).

opening between the vocal cords through which the air flows is termed the glottis. More extensive details of the anatomy of the larynx are presented in Chapter 4 of Hardcastle (1976).

During normal breathing, the vocal cords are relaxed, and the glottis is fully open, so that the air flow encounters little obstruction. However, when the appropriate muscles contract, the glottis closes, and the air flow is obstructed. This causes a differential air pressure to build up across the glottis. When sufficient tension is put on the vocal cords, the glottis becomes closed to such a degree that the air flow may stop. This causes the pressure differential to increase and force the glottis open again. The sudden puff of air that is emitted both reduces the pressure differential and produces a force (by means of the Bernoulli effect) which pulls the vocal cords back together again (van den Berg, 1968). Thus, as long as both the air flow and the tension on the vocal cords are maintained, a periodic series of pressure pulses is emitted from the larynx. These puffs of air are termed *glottal pulses* and the sounds produced with this type of excitation are termed *voiced* (§2.1.4). This brief description of how sound is produced by the vocal cords is expanded in §2.3.1.1 where I discuss the models which are invoked to describe their operation. Further details of the various physiological factors involved in the operation of the larynx are given by Lieberman and Blumstein (1988, Chapter 6) and Hardcastle (1976, Chapter 4).

2.2.1.3 The supra-laryngeal vocal apparatus

The primary functions of the supra-laryngeal vocal apparatus are to aid in providing air and food to the body. However, in most mammals, and especially humans, they are adapted to help produce different types of sound (Lieberman, 1975). The vocal purpose of the supra-laryngeal vocal tract (in the remainder of this thesis I refer to it simply as the *vocal tract*) is to modify the character of the sound coming from the larynx, and radiate it into the air.

The major components of the vocal tract are shown in the cross-sectional view displayed in Fig.2.3. The important characteristic of the vocal tract for speech production is that it can attain a wide variety of shapes, by movement of the *articulators*. These are the tongue, lips, velum and jaw. Different configurations of vocal tract shape modify the sound character in different ways (see §2.3.1.3), and thus sequences of dif-

ferent sounds can be produced by moving the articulators in various ways. The vocal tract can also be configured so that it constricts the flow of air to such an extent that the resulting turbulence produces a sound. Sounds produced by such turbulent excitation are termed *unvoiced* (§2.1.4. Speech sounds may also be produced by a *mixed excitation* containing both voiced and unvoiced components (e.g. for the sound /z/).

The exact vocal tract configuration by which different sounds are produced is beyond the scope of this thesis. Many references are available, both in the field of speech-language therapy and the field of physiological phonetics, which describe the vocal tract and the different ways in which people configure it to produce speech sounds (cf. Fant, 1960; Hardcastle, 1976, Chapter 6; Lieberman and Blumstein, 1988, pp114–131; Skinner and Shelton, 1978, Chapter 8). It is relevant, however, to note that the vocal tract can assume only a finite range of shapes, which limits the types of sounds that can be produced (Lieberman, 1975, pp68–74). The rate at which the vocal tract can alter its shape is also finite, which places a limit on how rapidly a person can talk.

2.2.2 Speech perception

2.2.2.1 Physiology of the ear

A cross-sectional view of the ear is shown in Fig.2.5. The ear is divided neatly into three parts, termed the outer, middle and inner ears.

The outer ear channels the sounds from outside the head to the eardrum, causing it to vibrate. The channel has a fairly broad resonance centred at 3kHz, which tends to amplify sounds in the range of 2–6kHz (Kinsler *et al.*, 1982, p257). The eardrum is a sealed membrane which forms the boundary with the middle ear.

The middle ear acts as a mechanical impedance matching transformer. It also protects the inner ear from damage due to high intensity sounds. The mechanical vibrations of the eardrum impinge on a lever type arrangement of three small bones (the ossicular chain, consisting of the malleus, incus and stapes), which transfers the vibrations to another membrane, termed the oval window. This is the boundary with the inner ear, and has an area about 17 times less than that of the eardrum. A transformation between the high impedance of the outer ear and the low impedance of the inner ear is effected by this area ratio, in conjunction with the lever arrangement of the ossicular chain (Kolston, 1989).

The ear has two built-in defense mechanisms against very high intensity sounds. Firstly, the mode of vibration of the stapes depends upon the sound intensity. For high intensity sounds, it vibrates in a way that is less efficient at transferring energy to the inner ear. Secondly, reflexive muscles in the inner ear act on the stapes and malleus when high intensity sounds occur, pulling the stapes from the oval window and hence reducing the amount of energy transferred through to the inner ear (Flanagan, 1972, §4.12).

The inner ear, or *cochlear*, consists of a spiral, fluid filled tube, divided longitudinally into two cavities by the *basilar membrane* (Fig.2.5*b*). The cochlear acts as a transducer, converting mechanical motion at the stapes into nerve pulses which are sent to the brain.

The vibration of the stapes on the oval window induces pressure waves in the fluid, which move along the upper cavity of the cochlear. Energy is transmitted to the lower cavity by movement of the basilar membrane or directly through a small hole at the apical (furthest away) end of the cochlear. As the pressure wave moves along the upper cavity of the cochlear, it displaces the basilar membrane. However, each portion of the membrane is most responsive to a particular range of frequencies, and so the position along the basilar membrane where the maximum amplitude response occurs

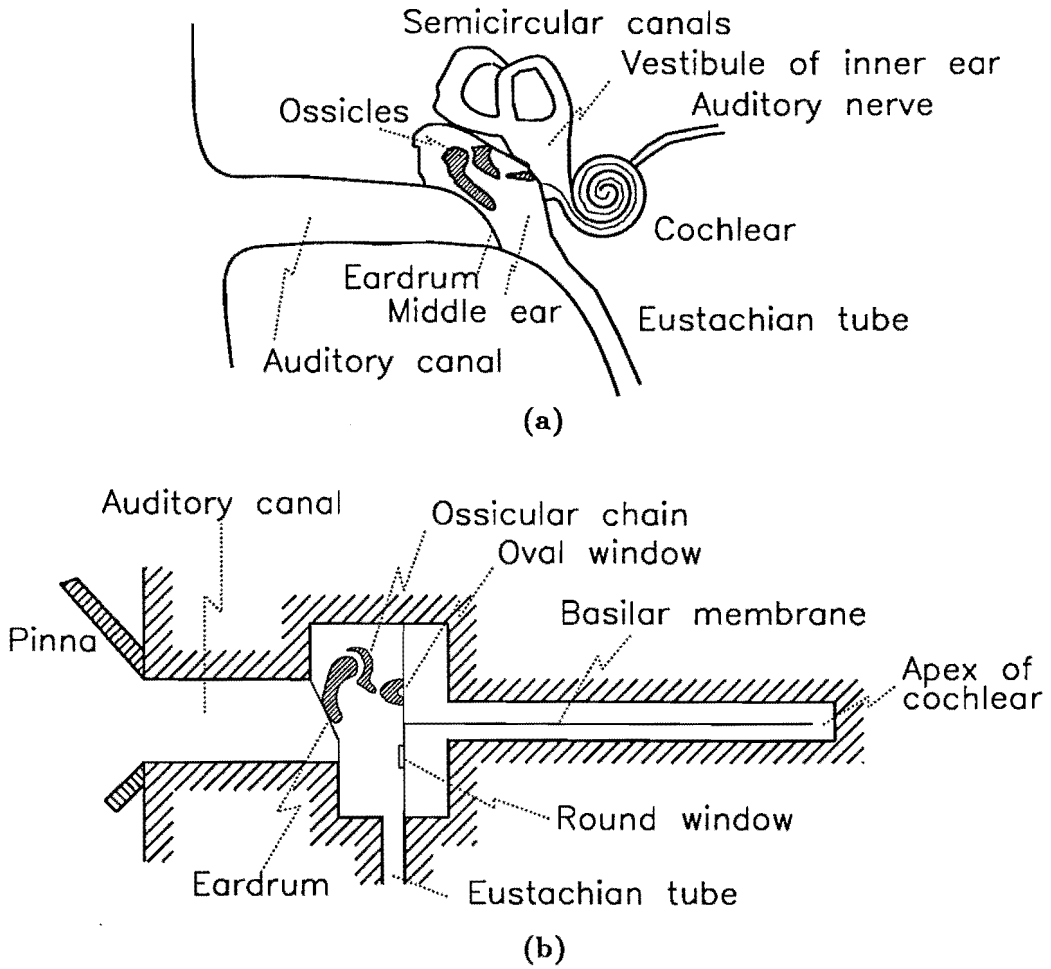


Figure 2.5. **a:** Cross-sectional view of the human ear, and **b:** diagramatic view of the components of the ear (after Skinner and Shelton, 1978).

indicates the frequency of the sound wave (Greenberg, 1988). The response peak ranges from about 20kHz at the basal (next to the middle ear) end of the membrane, to about 20Hz at the apical end (Kinsler *et al.*, 1982, Chapter 11).

Conversion of the mechanical motion of the basilar membrane into nerve pulses is performed by many thousands of hairs and associated nerve receptors. The hairs are bent by the motion of the membrane, and thereby produce electrical signals. These are then converted into nerve pulses which are transmitted to the brain (Flanagan, 1972, §4.14).

Further details of the physiology of the ear, as it relates to hearing, are given by Flanagan (1972, chapt4), Kinsler *et al.* (1982, Chapter 11) and Kolston (1989, Chapter 2).

2.2.2.2 Perception of sounds by humans

The physiological structure of the ear, as outlined in §2.2.2.1, has important consequences for the way in which sounds are perceived by humans. In this section, I discuss some of the characteristics of sound perception that are pertinent to speech perception.

The first characteristic of human hearing of interest here is that the sensitivity of the ear to sounds varies according to the frequency of the sound. Maximum sensitivity

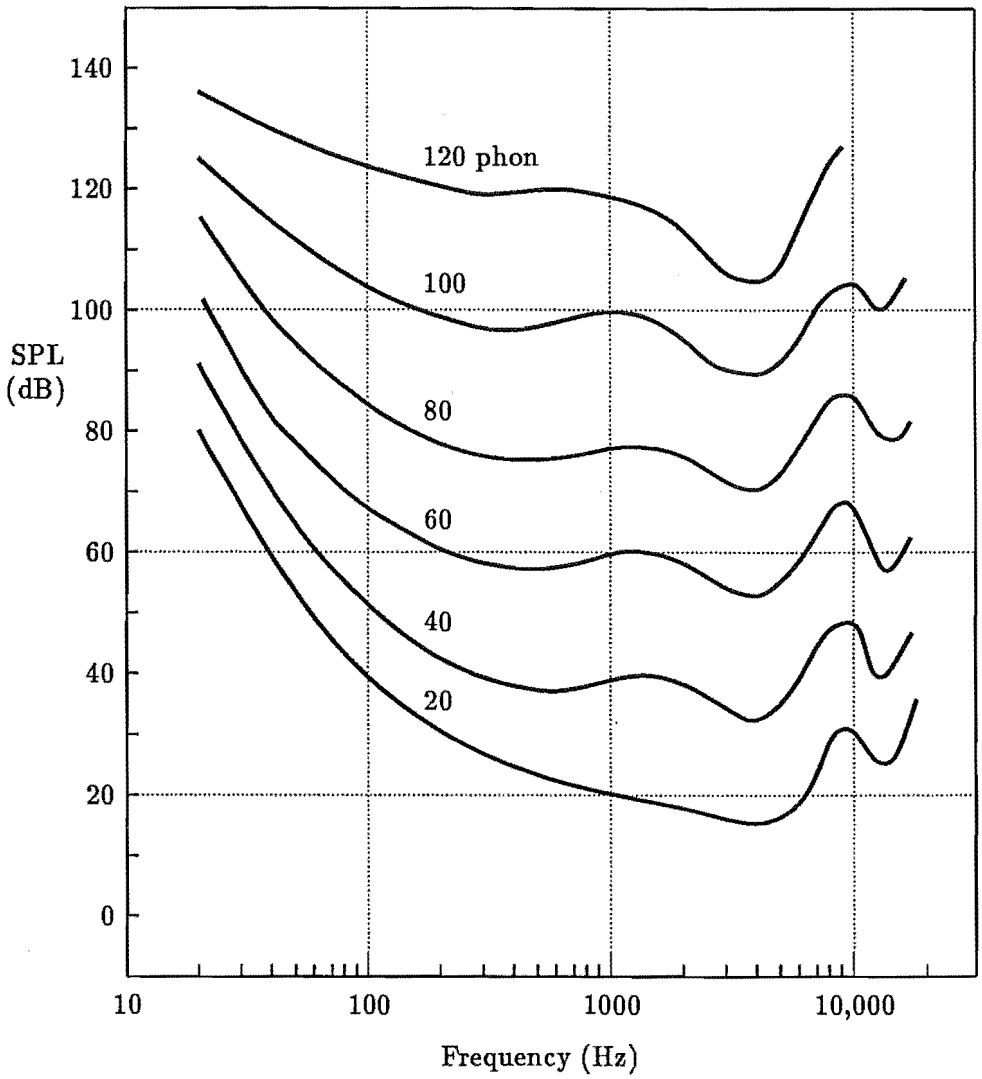


Figure 2.6. Graph illustrating the characteristics of loudness perception by humans. "Equal loudness' curves indicate the intensity which sounds of different frequencies must possess in order to be perceived as having the same loudness (after Kinsler *et al.*, 1982).

occurs for sounds in the range 1 to 3kHz (Skinner and Shelton, 1978, p128). The perception of the relative loudness between two sounds is not linearly related to the measurable intensity of the sounds. In addition, sounds of different frequencies that are perceived as being of equal loudness have different intensities (Fig.2.6).

A second characteristic of sound perception (especially of periodic signals) is that of *pitch*, which is the term given to the perception of a sound's fundamental frequency. However, the perceived pitch is not a linear function of frequency (in Hertz). A frequency scale which is related to the perception of pitch by humans is the mel scale, shown in Fig.2.7 (Kinsler *et al.*, 1982, p273). A further interesting characteristic of pitch perception is that the fundamental need not be present (cf. Houtsma and Goldstein, 1972).

The cochlear can be imagined as acting like a bank of overlapping filters. The width of the frequency bands manifests itself in two ways. Firstly, two sinusoidal tones must be separated in frequency by some minimum amount before they are perceived separately (Lieberman and Blumstein, 1988, pp159–161). Secondly, a sinusoidal tone

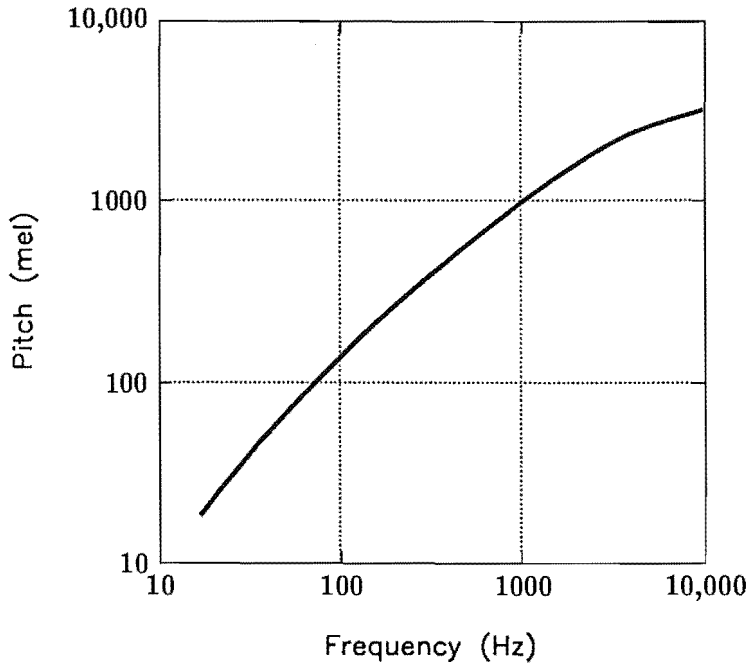


Figure 2.7. The perceptual, or mel, scale of pitch versus actual frequency of sound (after Kinsler *et al.*, 1982).

is masked out by noise (of equal intensity) only if the noise has a bandwidth greater than some *critical bandwidth.* These critical bands are roughly 1/3rd of an octave in width. They probably relate to physical characteristics of the cochlear (Kinsler *et al.*, 1982, §11.7). They are often termed *Bark bands* in the literature (cf. Traunmüller and Branderud, 1989).

The masking of sounds by other sounds mentioned in the previous paragraph also extends to sounds outside the critical band if the interfering sound is of sufficient intensity. An interfering sound usually masks out sounds of higher frequency that are more than about 20dB lower in intensity (cf. Skinner and Shelton, 1978, p136; Atal and Schroeder, 1979).

Masking only occurs when sounds are presented simultaneously. If different tones are presented sequentially, listeners are able to detect frequency differences of the order of 1/30th of the frequency of the tone, and amplitude differences of the order of 3dB (Flanagan, 1972, pp279–286; Kinsler *et al.*, 1982, pp264–166).

Because the cochlear detects sounds according to their short-term spectral magnitude, it is insensitive to phase distortions of the sound waveform (Schroeder, 1975). Of course, it is sensitive to phase distortion that is sufficiently severe that the waveform is distorted over an interval longer than the effective time-constant of each band-pass filter (cf. Schroeder, 1983).

Another characteristic of hearing is a consequence of the non-linearities introduced by both the cochlear and neural processing of sounds. This is the generation of sum and difference frequencies when two or more sounds of different frequencies are present. Everyone is familiar with the "beating" produced when two tones close ($<$ 10Hz difference) in frequency are simultaneously played. This effect is due to non-linearities in the cochlear, and vanishes when the two tones are separately, but simultaneously, applied to opposite ears. However, for tones that are separated by more than

a critical band, sum and difference frequencies are produced by non-linearities in the neural processing of the sounds (they are present even when the tones are presented separately, but simultaneously, one to each ear). These sum and difference frequencies allow a fuller appreciation of the subtleties of a musical harmonic scale. They also enable us to regenerate within our brains any "missing" fundamental frequency of harmonically related sounds. This last characteristic is helpful for the success of limited bandwidth speech reproduction schemes, which often discard the fundamental frequency, but without apparent degradation of the perceived sound.

2.2.2.3 Psychoacoustic characteristics of human speech perception

In addition to the characteristics of human sound perception discussed in §2.2.2.2, the perception of speech has characteristics which indicate that specialised neural processes are employed (Stevens and House, 1972). When speech sounds are being listened to, electrical activity in the brain is localised to a greater extent than it is for musical sounds (McAdam and Whitaker, 1971). Various psychoacoustic experiments have indicated that people employ different types of neural processing for speech than for other sounds. For instance, in experiments with synthesised speech-like sounds, listeners who were told to expect speech sounds could clearly understand words in the sounds, while listeners who were not expecting speech sounds reported them as "buzzes", "whistles" or other non-speech sounds (Remez *et al.*, 1981). Other studies (cf. Cutting, 1974) have suggested that the brain contains "property detectors" which respond to certain types of signals (e.g. frequency transitions).

Another characteristic of speech perception is that speech sounds often appear to be perceived in a categorical, rather than discriminatory, fashion (Cutting and Rosner, 1974; Darwin, 1976, pp206–217; Lieberman and Blumstein, 1988, pp152–159; Stevens, 1989). By varying a single acoustic feature in a synthetic speech generator (for example, the voicing delay time for stop consonants), a range of sounds can be produced which encompass several "real" speech sounds as well as the "speech-like" sounds which lie between them on the "feature continuum". When people are asked to identify sounds from such a continuum, their classification exhibits sharp transitions between categories. In addition, people cannot discriminate between two sounds if they are both members of the same category, although they can easily discriminate between sounds that are members of different categories, even when the acoustic difference may actually be less than in the first case. It appears that the brain assigns a sound to a particular category (usually equated to a phoneme) and then "forgets" the actual sound (Cutting and Rosner, 1974). This allows a great saving in the memory required for words and sentences, because only the category needs to be stored, not the actual acoustic pattern. The "category perception" mechanism is especially marked for consonant sounds (Lieberman and Blumstein, 1988, pp156–159). By contrast, perception of other sounds is usually more "continuous", in that smooth variation in a acoustic feature results in a smooth variation in the response.

2.3 Speech modelling

The information in speech can be described in terms of various phonetic features (§2.1.3.1). In order to extract these features from speech sounds using signal processing techniques, it is useful to have a model of how a typical speech sound is produced. Furthermore, in order for the features to be extracted accurately, so that they adequately embody the desired information, the model must be physically and anatomically realistic. Models on which the analysis of speech signals are based generally mimic, in

approximate ways, either the human speech mechanism or our understanding of how speech is perceived by humans. Models of the human speech production mechanism (§2.3.1) generate features that are related to the anatomy and aerodynamics of the vocal organs (e.g. the vocal tract shape and glottal source signal §2.2.1). Perceptual models (§2.3.2), however, are employed to analyse speech in terms of the auditory and neural characteristics of hearing that are considered to be important for speech perception (§2.2.2).

As well as serving as bases for the extraction of features from speech sounds, the models mentioned above can themselves be studied and refined to gain a greater understanding of the actual processes of speech production (cf. Ishizaka and Flanagan, 1972; Fant, 1986) and perception (cf. Kolston, 1989). Such studies are often accompanied by extensive physiological experimentation and measurements to confirm the validity of whatever model is being investigated (cf. Boves, 1984, Chapter 4; Sondhi, 1979).

2.3.1 Speech production models

To be useful, any model of speech production must be based on the physical processes by which the vocal apparatus generates speech (§2.2.1). The glottal voice source is usually modelled as a self-oscillating mechanical/aerodynamic oscillator (Flanagan, 1972, pp246–251), while unvoiced sounds are modelled as being caused by air turbulence at a constriction in the vocal tract (Flanagan, 1972, pp251–259). The vocal tract is modelled as a tube of variable cross-section (Stevens and House, 1955), often represented by its electrical analogue, a transmission line (Fant, 1960; Flanagan, 1972). Radiation of sound from the mouth is often modelled by considering the radiation from a small piston on the surface of a sphere (Fant, 1960).

In §2.3.1.1 and §2.3.1.2 I discuss some of the models which explain how sound energy is generated in the vocal apparatus. §2.3.1.3 describes the ways in which the vocal tract is modelled. Finally, §2.3.1.4 introduces the simplified "source-filter" model of speech production, which is the most popular model invoked by analysers of speech sounds.

2.3.1.1 The glottal excitation

The glottal excitation is a quasi-periodic series of air pulses emitted by the vibrating vocal cords (§2.2.1.2). Models of how this excitation is produced usually consist of a mechanical model of the vocal cords which interacts with an aerodynamic model of the air flow through the larynx. Fig.2.8 illustrates the "two-mass" model of the vocal cords developed by Ishizaka and Flanagan (1972) (See also Flanagan *et al.*, 1975). The vocal cords are modelled by two mass-and-spring oscillators. The spring stiffness and damping factors represent the properties of the vocal cord tissue and the tension which is put on them during voiced speech. These parameters can be determined by physiological measurements (van den Berg *et al.*, 1957). Two masses are employed to simulate the measured characteristics of vocal cord vibration more closely than is possible with only a single mass (Ishizaka and Flanagan, 1972; Gupta *et al.*, 1973). Details of the acoustic signals produced by several modifications of the basic two-mass model are described by Koizumi *et al.* (1987).

An alternative mechanical model of the vocal cords is a single mass model which opens like a "zipper" to avoid the sudden contact–no-contact discontinuity which occurs in the two-mass model (Childers *et al.*, 1986). Some observations of vocal vibration made with high-speed cinegraphy show the cords performing in this manner. A model that includes a parallel "leak-opening" to the two-mass model has been proposed by

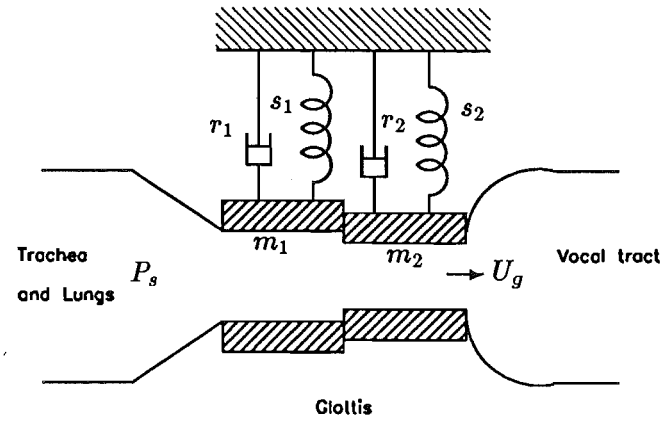


Figure 2.8. Simplified two-mass model of vocal cord vibration (after Ishizaka and Flanagan, 1972). P_s is the sub-glottal pressure, U_g the glottal flow velocity, while r_i, s_i, and m_i, $i = 1, 2$ represent the equivalent viscous resistances, elastances, and masses respectively of the two mass components of the vocal cords.

Cranen and Boves (1986) in order to account for leakage of air past incompletely closed vocal cords.

The mechanical models of the vocal cords express the variation in area of the glottal opening, and hence the flow of air through the glottis (Ishizaka, 1981). Feedback from the airflow to the mass-spring model arises from the forces exerted on the cords by the air pressure (both static and that caused by the Bernoulli effect). The air flow is also determined by the pressure difference across the glottis, which is the difference between the (effectively constant) lung pressure and the instantaneous pressure at the start of the vocal tract. The impedance which the vocal tract manifests to the larynx (and hence the pressure at that point) varies with time, both because of the varying shape of the tract (Bickley and Stevens, 1986), and because of pressure variations caused by resonances in the tract (Koizumi *et al.*, 1985).

The equations which define the mechanical-aerodynamic models of glottal excitation require a great deal of iterative computation to solve. Hence these models are seldom used to analyse or synthesise speech (A speech synthesiser employing the two-mass model (Sondhi and Schroeter, 1987) requires more computing power than that available in a Cray-1 supercomputer in order to produce speech in real-time).

By modelling the glottal excitation as independent of the vocal tract, the glottal flow can be calculated without laborious iterations. However, ignoring the influence of the vocal tract on the glottal flow introduces some errors which may be significant for certain sounds. The influence of the vocal tract arises mainly from the pressure variations caused either by the first formant resonance (since that is close in frequency to the pitch frequency, Ananthapadmanabha and Fant, 1982) or by sudden changes in tract shape (such as the closure of the tract during a stop sound, Bickley and Stevens, 1986). These influences tend to "skew" the glottal flow so that it is not simply proportional to the glottal area.

Ananthapadmanabha and Fant (1982) propose a formulation of the "true" glottal flow which is a combination of a (short-circuit) source component together with a "ripple" component caused (largely) by the effects of the first formant. In a similar vein, Rothenberg (1981,1983) parametrises the "shape" of the glottal area and then uses a simple LC model of the first formant to calculate the actual glottal air flow. Titze *et al.* (1983) model the glottal vibration by a combination of two sinusoidal components,

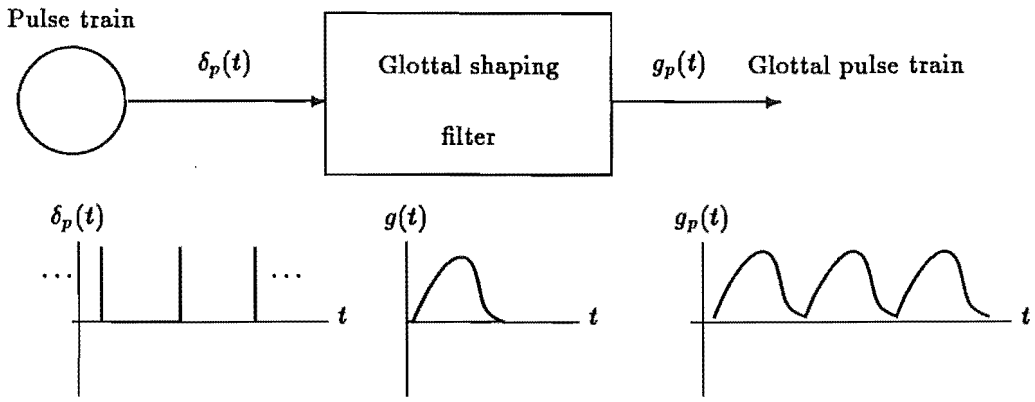


Figure 2.9. Simple model of the glottal excitation. The impulses are spaced by the pitch interval. The glottal shaping filter models the "shape" of each glottal pulse

characterising the upper and lower portions of the vocal cords. The glottal area, contact area, and hence glottal flow waveforms are described by simple combinations of the two components. Comparisons between glottographic waveforms (cf. §3.4.1) obtained from real speakers and the waveforms predicted by the model indicate a close match, although the model does not explicitly explain the effects of vocal tract interaction (Titze *et al.*, 1983).

The simplest model of glottal excitation is a quasi-periodic train of impulses, each filtered by a glottal shaping filter (Fig.2.9). In this model, the impulses are spaced by the pitch interval to accord with the quasi-periodicity of the glottal excitation. The filter impulse response models the shape of each pulse of air, which may vary with each pulse (Flanagan, 1972, p233). Often simple shapes such as triangular, sine-squared or other approximations are used as the "filter" (Rosenberg, 1971; Flanagan, 1972, pp232–246), since these are simple to (digitally) implement. In addition, they approximate the average spectral characteristics of the actual glottal excitation (a -12dB/octave slope on the magnitudes of the spectral coefficients, cf. Sundberg and Gauffin, 1979) reasonably well (Flanagan, 1972; Holmes, 1973).

The independent glottal excitation model (see also §2.3.1.4) is often employed in the analysis of speech because the assumptions of linearity and independence between the glottal and vocal tract models simplify the analysis (cf. §3.2,§3.4).

2.3.1.2 Unvoiced sound source

Unvoiced sounds are caused by turbulence in the air flow through the vocal apparatus. The simplest model for unvoiced sounds is a white noise source, which replaces the glottal excitation source. However, this assumes that speech sounds are composed of exclusively voiced and unvoiced segments, which is obviously not true (§2.1.4.2). This drawback can be overcome by providing for a "mixed excitation" sound type, in which both sound sources are present to some degree. When analysing speech sounds, however, it is often difficult to determine the correct "mixture" of voiced and unvoiced excitation.

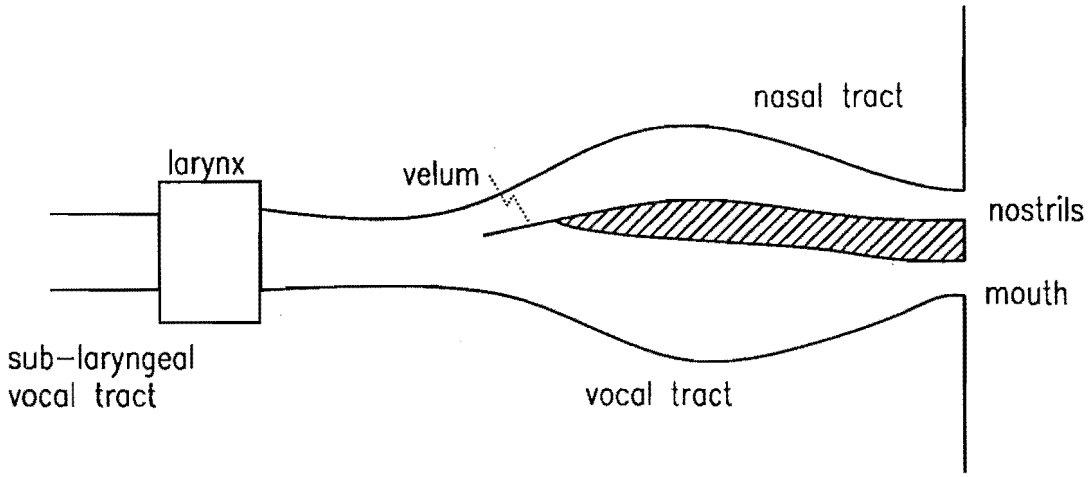


Figure 2.10. Acoustic model of the vocal tract, as a tube of varying cross-section (after Flanagan *et al.*, 1975).

Since unvoiced sounds are produced at constrictions in the vocal tract, one can examine the cross-sectional area of the tract (§2.3.1.3) and add some noise to the excitation if the area is small enough that turbulence is to be expected (Flanagan, 1972, p251). In addition, the noise can be added to the vocal tract at the point of constriction, which is more realistic than simply assuming that it originates from the glottis (Sondhi and Schroeter, 1987; Flanagan *et al.*, 1980).

2.3.1.3 Vocal tract models

The vocal tract consists of the mouth and nasal passages (§2.2.1.3). By virtue of the wide range of shapes which it can assume, it is able to affect the character of the sounds propagating through it in a great variety of ways. The effect of the vocal tract can be evaluated by considering the transmission of acoustic sounds through a tube of varying cross-section (Fig.2.10; Flanagan, 1972, Chapter 3; Flanagan *et al.*, 1975). Radiation from the ends of the tubes (at the mouth and nose) is modelled by considering the characteristics of a vibrating piston in a spherical baffle (Fant, 1960, pp34–36). At the glottis, the tube is terminated by a varying size opening. The velum acts like a valve, connecting or disconnecting the nasal passages from the rest of the tract.

The dimensions of the vocal tract are smaller than the dimensions of a wavelength for sounds of frequencies less than about 4kHz. Thus the wave propagation through the tract can be considered to be effectively planar (Sondhi, 1974). In addition, the walls of the vocal tract can be assumed to be effectively rigid for sound frequencies greater than 500Hz (Sondhi, 1974). By taking into account these approximations, the vocal tract can be modelled in one dimension as a transmission line, with a side branch for the nasal passages (Fig.2.11). The transmission line can usually be modelled usefully by a lumped-parameter approximation of about four stages (Flanagan, 1972, pp80–83), although more stages are obviously more accurate. The termination of the transmission line at the mouth and nose is modelled by an impedance calculated from the characteristics of a radiating piston (Fant, 1960, pp35–36,pp61–63). The glottis can be considered as a time-varying impedance.

The acoustic tube model of Fig.2.10 can also be approximated by a series of uniform tubes of different diameters (Fig.2.12). Often the nasal tract is also neglected, since this simplifies the mathematical analysis. Nasal sounds can usually be approx-

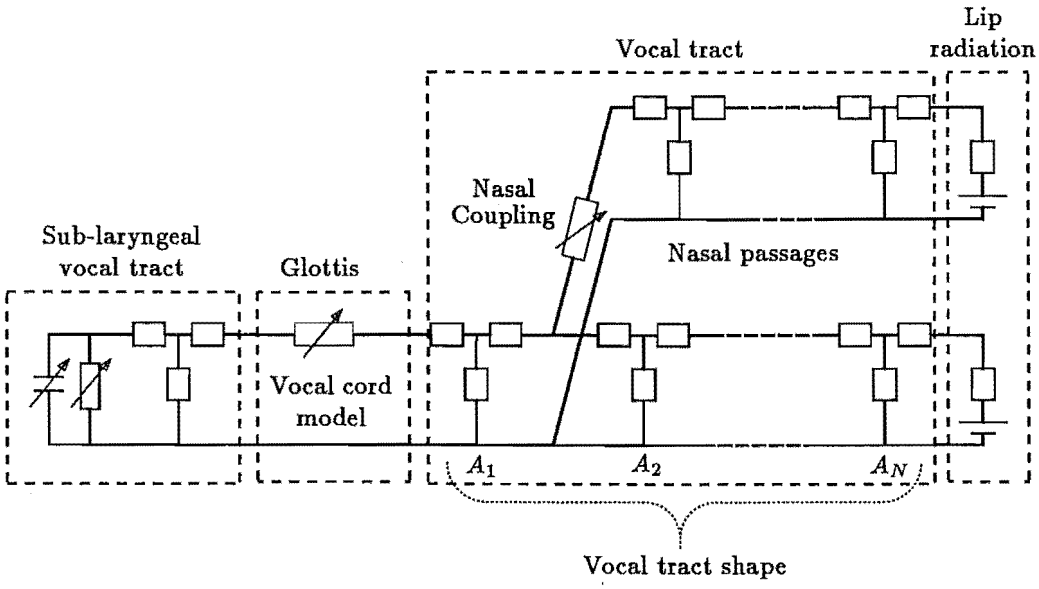


Figure 2.11. Transmission line model of the vocal tract. The transmission line is represented by a lumped-parameter approximation (after Flanagan *et al.*, 1975).

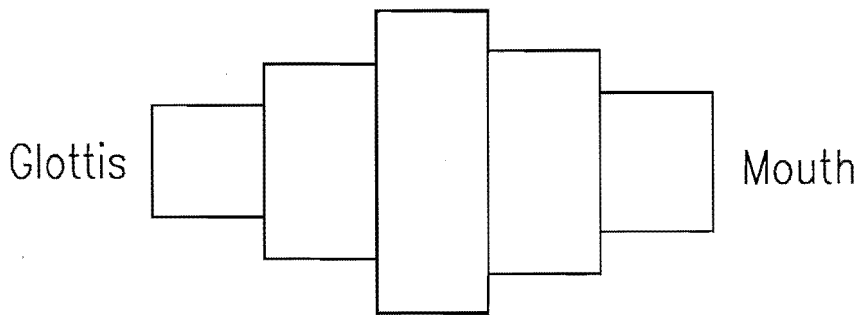


Figure 2.12. Model of the vocal tract as a series of uniform tubes. The nasal tract is ignored in this representation.

imated by additional sections in the uniform tube model. If losses are ignored, the response of the uniform tube model can be calculated by considering the reflections caused at each discontinuity. Such an analysis leads to a set of *reflection coefficients* which characterise the vocal tract (see §3.2 for details of an analysis technique which makes use of this representation). In addition, this model is equivalent to the lumped-component transmission line discussed above.

For most speech sounds, the vocal tract is configured as one or two cavities. Hence the shape of the vocal tract (tube) can be parametrised by considering the position and amount of the constriction and the sizes of the cavities (Flanagan *et al.*, 1980). Flanagan *et al.* (1980) employ this method of encoding the *area function* of the vocal tract in their articulatory speech synthesiser. By treating each section (e.g. glottis-to-velum, velum-to-constriction etc.) separately, Sondhi and Schroeter (1987) construct input-output matrices for each, and then simply multiply them together in the frequency domain to obtain the response for the entire vocal tract (including the nasal passages). Mermelstein (1973) presents a comprehensive model of the physical structures of the vocal tract which allows an investigation into the relationships between

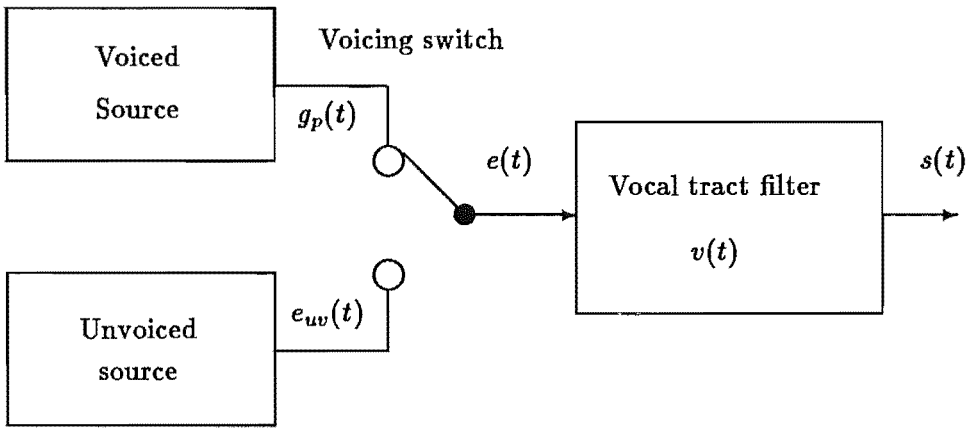


Figure 2.13. Source-filter model of the speech production process. The source is modelled by a quasi-periodic train of pulse (voiced sounds) or a sequence of random noise (unvoiced sounds). The vocal tract filter is a time-varying filter which also includes the effect of the lip radiation.

particular speech sounds and their corresponding articulatory parameters.

2.3.1.4 The source-filter model of speech production

The model of speech production most commonly employed in speech analysis schemes is the *source-filter* model (Fant, 1960), shown in Fig.2.13. In this model the source and vocal tract are modelled independently, which greatly simplifies (by comparison with interactive models) the extraction of features from speech signals.

In the source-filter model, the vocal tract is represented by a time-varying filter, which also includes the effect of radiation from the lips. The source is modelled by a train of pulses for voiced sounds (§2.2.1.2) and by white noise for unvoiced sounds (§2.2.1.3).

In this thesis, I use the symbols $e(t)$, $v(t)$ and $s(t)$ to refer to the source (or excitation) signal, vocal tract impulse response and resulting speech signal respectively. When discussing voiced speech signals, I use the symbol $g(t)$ to denote the glottal excitation. The sampled version of each of these quantities is denoted by $e[n]$, $v[n]$, $s[n]$ and $g[n]$ respectively.

For short segments of speech (a few pitch periods), the sound emitted by the lips in the source-filter model is described by the following convolutional equation:

$$s(t) = e(t) \odot v(t) + c(t) \tag{2.1}$$

where I introduce the symbol $c(t)$ to represent any part of the speech signal that cannot be represented by the convolution. In keeping with a variety of developments in this laboratory (cf. Bates, 1982; Bates and McDonnell, 1986; Brieseman *et al.*, 1987), $c(t)$ is called the *contamination*, and includes the effects of non-linearities, source-filter interaction and additive noise. Note that the convolution (2.1) is only true for short segments of speech because the vocal tract impulse response changes as different sounds are uttered. A formulation of (2.1) which explicitly takes this variation into account is (Brieseman *et al.*, 1987)

$$s(t) = \sum_{m=1}^{M} v_m(t) \odot g_m(t - \tau_m) + c_m(t) \tag{2.2}$$

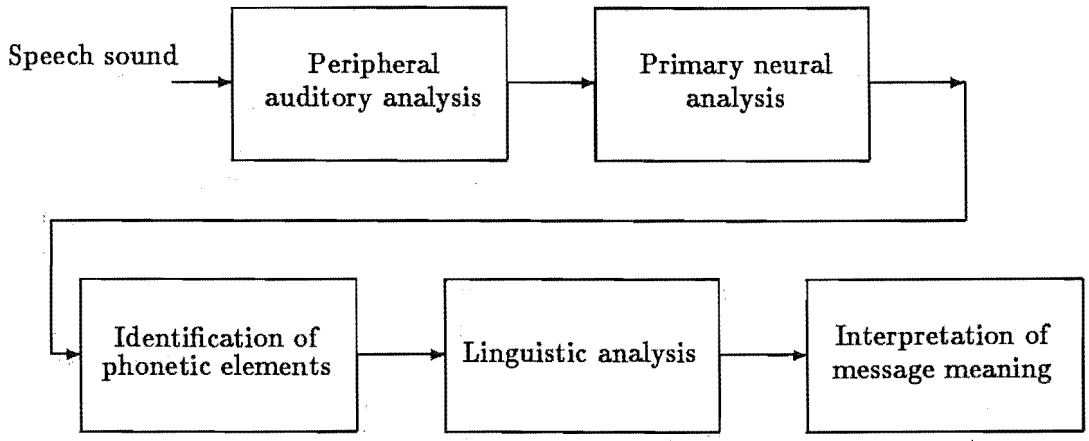


Figure 2.14. Block diagram of the stages in the perception of speech sounds (after Fant, 1973, Chapter 10.)

where the speech signal is assumed to be divided into M individual segments, each of which is characterised by a different tract response $v_m(t)$, glottal pulse $g_m(t)$ and contamination $c_m(t)$. The m^{th} glottal pulse $g_m(t)$ occurs at the instant τ_m. Any part of the speech signal in the m^{th} interval which is due to overlap from the signals in the $(m-1)^{\text{th}}$ interval is accounted for by the contamination term $c_m(t)$.

By employing this linear model of speech production, signal processing techniques such as deconvolution and Fourier analysis can be straightforwardly applied to extract features from the speech signal.

2.3.2 Models of speech perception

Perceptual models are based on the way in which sounds are perceived by humans (cf. Fant, 1973, Chapter 10). They involve characterising the speech signal in terms of its perceptually important features (§2.2.2). By doing this, they attempt to overcome some of the difficulties, outlined in §2.1.3 and §2.1.4, of the wide variability in the acoustic representation of phonemes (cf. Syrdal and Gopal, 1986). The model of speech perception that is commonly invoked consists of several stages, each one of which transforms the speech signal into a representation with less redundancy. Figure 2.14 depicts the stages in one model of speech perception (Fant, 1973, Chapter 10).

The first stage of speech perception is the transformation of sounds into nerve pulses, which occurs in the cochlear (§2.2.2.1). The response of the cochlear can be modelled by a bank of filters (Allen, 1985; Kolston, 1989). Usually this filter-bank is implemented by performing some standard spectral analysis of the sound (§1.3.1) and then modifying the spectral coefficients to conform with the experimentally obtained characteristics of the cochlear (cf. Allen, 1985; Syrdal and Gopal, 1986). The perceptual characteristics which are employed to modify the acoustic features are those of relative loudness perception, critical bandwidths and relative pitch (see §2.2.2.2).

The neural processes by which sounds (whether speech or not) are perceived by humans are still not fully understood, but their characteristics can be partially inferred from psychoacoustic experiments (§2.2.2.2 and §2.2.2.3). The observed characteristics can then be accounted for by suitable transformations of the acoustic features of the

sounds. Comprehensive models of perceptual processes are yet to be developed, but approaches based on neural networks (Lippmann, 1987) may eventually prove useful. Neural networks have also been employed to model the process of sound-to-neural transduction that occurs in the cochlear (Lyon and Mead, 1988).

Another approach to modelling the phonetic/linguistic levels of the perceptual model shown in Fig.2.14 is to describe the linguistic units of a language in terms of distinctive acoustic features (Jakobson *et al.*, 1961). This approach is appealing because of evidence (§2.2.2.3) that people perceive speech in terms of distinct phonemes (Stevens, 1989). Unfortunately, the acoustic manifestations of the acoustic features of particular phonemes vary widely (§2.1.3.1), and work is still proceeding on ways to account for this variation (cf. Traunmüller and Branderud, 1989).

A model of human speech perception that is currently the subject of much debate (cf. Fowler, 1986; Remez, 1986 and other articles in the same journal issue) is the "motor theory", which postulates that humans perceive speech by determining the articulatory configuration required to make each type of sound heard. They then match that particular articulatory configuration to its associated linguistic "label" (cf. Tobias, 1972, pp47–56; Fowler, 1986; Lieberman and Blumstein, 1988, pp147–149). This "analysis-by-synthesis" model is supported by observations that some parts of the brain are common to both speech production and perception. However, the difficulties inherent in determining the vocal tract configuration from the speech sound only (Sondhi, 1979,1984; Bonder, 1983), cast doubts on how far the motor theory can be applied in explaining speech perception.

In conclusion, the most useful application of perceptual models seems to be as a method of modifying the features obtained by techniques, such as short-term spectral analysis, so that they relate more closely to the manner in which humans perceive those features. Evidence from several studies indicates that this approach improves the performance of word and phoneme recognition schemes (cf. Syrdal and Gopal, 1986; Cohen, 1989), especially speaker independent speech recognition (Bladon, 1985; Johns-Lewis, 1986).

Chapter 3

Established speech analysis techniques

This chapter describes some of the established techniques that are employed for analysing speech signals. Because the techniques that I introduce in subsequent chapters relate to glottal pulse estimation (Chapter 4) and low data rate speech encoding (Chapter 5), the thrust of this chapter is towards these areas.

In §3.1 through §3.3 I introduce the techniques by which the features of speech sounds are extracted. §3.1 describes techniques that extract features related to the prosodic structure of speech, §3.2 describes the linear prediction method of speech analysis, and §3.3 explains how speech can be analysed in terms of its spectral content. In §3.4 I describe the application of the above-mentioned techniques to estimating the glottal excitation waveform, while §3.5 is concerned with techniques for encoding speech at low data rates. §3.6 briefly introduces the application of speech analysis techniques to the problems of speech and speaker recognition, text-to-speech conversion, and the diagnosis of, and therapy for, speech disorders.

The speech processing literature is very extensive, and more details of the material introduced in this chapter can be found in the references cited. Some useful textbooks are those by Rabiner and Schafer (1978), especially with regard to the time domain, frequency domain and low data rate analysis techniques, Markel and Gray (1976), who comprehensively cover the various LPC techniques, Fallside and Woods (1985), who provide a wide background, especially on speech recognition topics, and Flanagan (1972), who covers many of the techniques of speech analysis and synthesis, together with their application to low data rate speech encoding.

3.1 Prosodic feature analysis

Prosodic features of speech are time domain characteristics, such as the variation of pitch during an utterance and the syllabic structure. The latter includes the classification of speech segments as voiced or unvoiced (§2.2.1), segmentation of speech into words, the changes in loudness that occur for each syllable, and determination of the speech "rate".

3.1.1 Loudness of speech

The loudness of speech is a para-linguistic feature which characterises the syllabic structure of an utterance. It also carries non-linguistic information such as the emotional state of the speaker (§2.1.3.1). For applications such as automatic speech recognition, the analysis of loudness must take into account how loudness is perceived by humans.

For other applications, however, such as speech compression, this is unnecessary, provided that the original loudness levels are preserved in the resynthesised speech.

The usual measure of speech loudness is the short-term energy envelope of the speech signal, defined as

$$E[m] = \frac{1}{L} \sum_{i=m-L/2}^{m+L/2-1} s^2[i], \tag{3.1}$$

where L is the length of the *window* that delineates the segment of speech that $E[m]$ refers to. The length of the window determines the resolution with which changes in loudness are reflected in $E[m]$. (3.1) can also be viewed as the convolution between the squared speech signal and $w[n]$, the window function:

$$E[m] = \frac{1}{L} \sum_{-\infty}^{\infty} s^2[i] w[m-i]. \tag{3.2}$$

The window implied in (3.1) is then given by

$$w[m] = \begin{cases} 1 & -L/2 \leq m \leq L/2 - 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

The right hand side of (3.2) can be regarded as a filtering operation, implying that the loudness measure is equivalent to the output of an envelope detector. The envelope $E[m]$ is usually (Rabiner and Schafer, 1978, p124) calculated only at values of m spaced much further apart than the spacing between the samples of $s[i]$. Care must then be taken that the Fourier transform of $w[i]$ is small enough at frequencies greater than half the new sampling frequency (of $E[m]$) so that aliasing does not occur. The bandwidth of the smoothing window needs to be large enough that changes in the loudness envelope occurring between syllables are adequately sensed, but low enough that variations within a pitch interval are smoothed out. Bandwidths between 20 and 50Hz are commonly employed, depending on whether the overall syllabic structure of an utterance or the finer detail within a syllable is required. A window, of simple form, that is often employed for envelope calculation is the Hamming window (see §1.3.1.1).

A drawback of the energy envelope as defined in the above paragraph is its large dynamic range. The root-mean-square (RMS) envelope or the mean absolute envelope are alternatives which are often more useful because they have the same dimensionality as the original signal. Fig.3.1 shows an example of energy, RMS, and absolute envelopes for a short segment of speech.

3.1.2 Voiced/unvoiced decision analysis

Voiced/unvoiced (VUV) analysis entails deciding whether short segments of speech are of voiced, unvoiced, or mixed type (§2.2.1). Several approaches to implementing such decisions have been proposed (cf. Siegel, 1979; Knorr, 1979). In this section I briefly describe several disparate methods that illustrate the range of techniques employed.

VUV classification is often associated with *silence detection*, which is the detection of gaps between words. The main difficulty that must be overcome in order to reliably classify speech and "silence" is to detect the difference between a quiet fricative sound and background noise (De Souza, 1983). Silence detection techniques are invoked for isolated-word recognition, in order to isolate the individual words and identify their start and end points (cf. Savoji, 1989).

One approach to VUV classification is based on the differences in spectral content between voiced and unvoiced sounds. The energy in the voiced excitation is

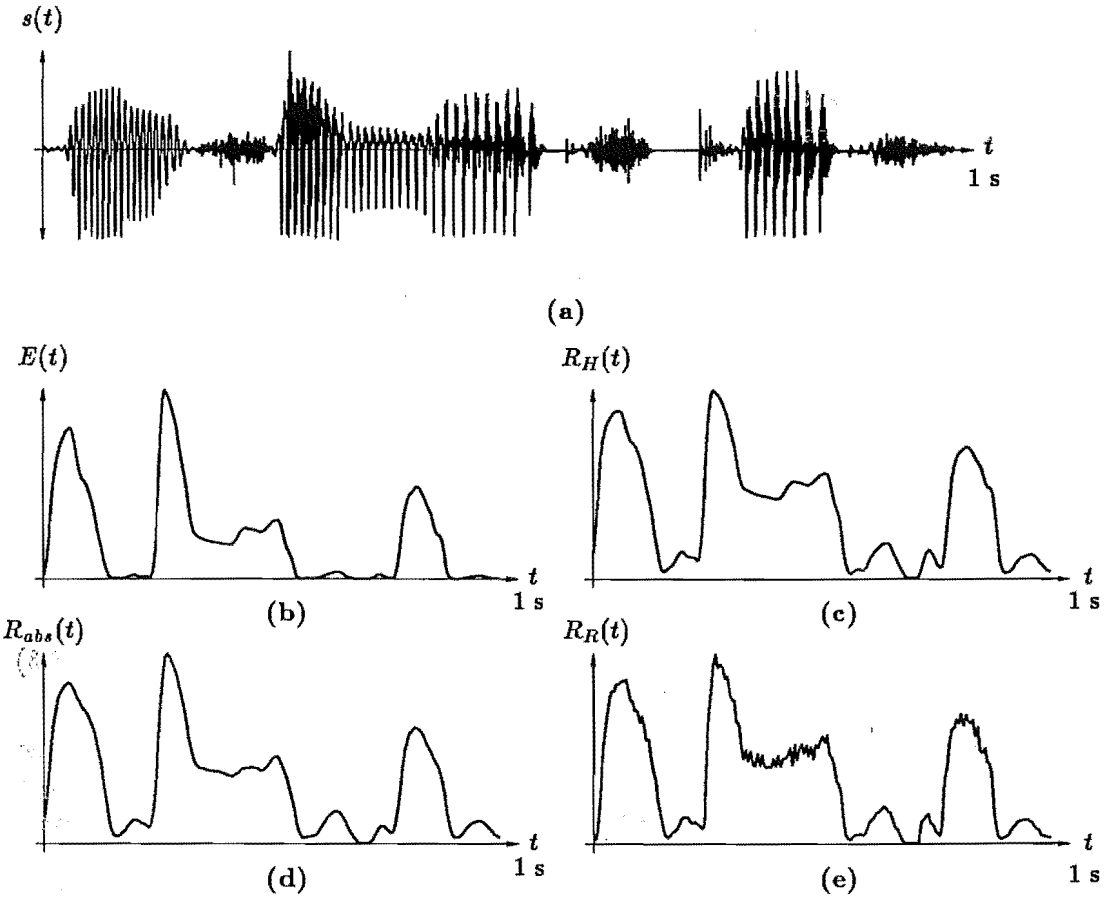


Figure 3.1. Example of a loudness contour for a speech signal. **a:** The speech signal. **b:** Energy signal, **c:** RMS signal, and **d:** absolute envelope signal. The envelope signals shown in *b,c* and *d* are calculated with a 40ms Hamming window. **e:** The energy envelope calculated with a 20ms rectangular window.

concentrated at low frequencies, while that for the unvoiced sounds occurs throughout a broader and higher band of frequencies (§2.1.4.2). Hence a decision based on comparison of the short-term energy in high frequency and low frequency sub-bands of the speech is an effective method of distinguishing between voiced and unvoiced sounds (Knorr, 1979). I employ a method similar to this when performing VUV analysis for the algorithms described in this thesis. The speech is low-pass filtered by a filter with a cutoff frequency of 200Hz. The RMS envelope $R_l[m]$ of this band is then calculated. In addition, the speech signal is differentiated to accentuate the high frequencies (this is simpler than performing a high-pass filtering operation). The RMS envelope $R_d[m]$ of the differentiated speech signal is also calculated. The RMS envelope $R_s[m]$ of the complete speech signal is computed in order to check for "silent" segments. The speech is classified according to the following rule:

Segment m is

$$\begin{aligned} &\text{Silent if} && R_s[m] < \xi_s \\ &\text{unvoiced if} && R_s[m] > \xi_s \text{ AND } R_l[m] < 1.25 R_d[m] \\ &\text{voiced if} && R_s[m] > \xi_s \text{ AND } R_l[m] > 1.25 R_d[m] \end{aligned}$$

The factor of 1.25 introduced above is required to ensure that the comparison between the two envelope signals produces reasonable results (Brieseman, 19XX).

An indication of the spectral content of a speech segment is also provided by the *zero-crossing rate*, which is defined as the number of times that the signal amplitude changes sign in that segment (Rabiner and Schafer, 1978, §4.3). A segment that contains mainly high frequency components has more zero-crossings than one that is comprised mainly of low frequency components (see Scarr, 1968). Hence the segment can be relatively easily classified as unvoiced, if its number of zero-crossings is above some threshold, or voiced, if this number is below the threshold. Usually, the energy of the segment is also examined, to decide whether the segment is of speech or silence (since silent segments still contain noise, and hence zero-crossings). This approach is useful for real-time applications because of its simplicity (Watson *et al.*, 1988).

Another method of making a VUV classification is to examine the results of a pitch analysis of the speech signal (§3.1.3). During segments of unvoiced speech, pitch analysis techniques usually fail to produce a pitch estimate. Hence the lack of periodicity can be used as an indication of whether the speech signal is voiced or unvoiced (cf. Gold and Rabiner, 1969).

A final VUV analysis technique is based on applying pattern recognition approaches to several features, usually including some of those mentioned above (Siegel, 1979; Siegel and Bessey, 1982). In this approach, the characteristics of voiced and unvoiced sounds are specified by "templates" of features. The templates are obtained by an analysis of many examples of the two types of speech sounds. Segments of speech are then classified according to the template they are closest to. This type of analysis can be easily incorporated into speech recognition schemes (§3.6.1).

3.1.3 Pitch detection

The *pitch*, or *fundamental frequency* of a voiced segment of speech is the reciprocal of the average interval between successive glottal pulses (§2.2.1.2). In this thesis I use the term *pitch period* to refer to the duration of a single period of glottal excitation. The term *pitch interval* identifies a segment of speech of one pitch period in duration. The pitch varies during an utterance as a result of non-linguistic and para-linguistic information in the speech signal (§2.1.3.1). For example, if the pitch rises towards the end of an utterance, it indicates that the utterance is a question. The purpose of pitch detection is to estimate all the individual pitch periods of the voiced parts of an utterance. The resulting *pitch estimates* specify the durations and instants of occurrence of each pitch interval in the utterance.

Accurate determination of the pitch periods of voiced speech is useful for several purposes. Firstly, the naturalness and intelligibility of the speech synthesised from several of the low data rate speech encoding schemes described in §3.5.2 is dependent on the accuracy of the pitch estimates. Secondly, certain techniques of speech analysis (see §3.2.2 and §3.3.3) are *pitch synchronous*, in that they operate on segments of speech aligned to each pitch interval. A third application for the use of the pitch estimates is that of diagnosing the medical state of the speaker's vocal cords (§3.6.4.1)

Methods for determining the pitch of a speech signal can be divided into those which operate on the speech in the time domain and those which operate on a frequency domain representation of the speech.

One time domain method of estimating the pitch of a speech signal involves calculating the autocorrelation of a short (20–50ms) segment of speech and examining its structure. Such an autocorrelation reveals the pitch period because it exhibits peaks at lags equal to the pitch interval (Rabiner and Schafer, 1978, §4.6). In some cases, however, especially when the first formant of the speech signal is relatively strong or of narrow bandwidth, the autocorrelation also contains peaks at lags corresponding to the formant frequency. Techniques such as centre-clipping (Sondhi, 1968), in which only

the tips of the few largest peaks are employed in the autocorrelation calculation, can help to reduce the effect that the formant structure has on the autocorrelation.

Other time domain techniques are based on pattern matching approaches (Gold and Rabiner, 1969; Tucker and Bates, 1978). Gold and Rabiner (1969) take measurements of the amplitudes of peaks in a segment of the speech waveform and then match sets of these measurements in various combinations in order to produce several estimates of the pitch period. The estimates are then compared and the pitch period for that segment is defined as the one that occurs in the majority of estimates. When the techniques that I discuss in other sections of this thesis call for a pitch analysis of the speech signal, I use a modification of the Gold and Rabiner technique (Tucker and Bates, 1978; Brieseman, 1984). The extrema of the speech signal are first located, with the instant of occurrence and amplitude of the m^{th} extrema denoted by t_m and A_m respectively. The pitch estimate at the instant t_m is obtained by finding the lowest value of P for which

$$|A_n - A_{n-P}| < \epsilon \qquad \text{for } n = m, m-1, \ldots, m-2P, \tag{3.4}$$

where ϵ is a threshold that determines the maximum "variability" that the peaks are allowed to possess between pitch intervals. ϵ is set to 20% of the magnitude of the largest peak in the current interval. The pitch period is given by $t_m - t_{m-P}$. This algorithm has been found to be useful for determining the pitch of musical sounds as well as speech sounds (Brieseman, 1984). In order to ensure that the pitch is estimated accurately, I filter the speech signal to remove the components with frequencies greater than 500Hz. In addition, the algorithm is limited to finding pitch estimates between 75Hz and 500Hz. These constraints are necessary to ensure that the formants do not adversely affect the pitch estimates (Brieseman, 1984). If children's voices are to be analysed, it is wise to increase both the higher limit and the filter cutoff frequency to about 900Hz (Watson, 1989).

Instead of estimating the pitch from the speech signal directly, it can be estimated from the residue that remains after removing the effects of the formant resonances (§3.2). Because the residue produced by filtering the speech signal with the inverse LPC filter often exhibits sharp spikes spaced by the pitch interval, this can be an effective method of pitch estimation. Descriptions of techniques that make use of the LPC residue are given by Markel (1972) and Markel and Gray (1976, Chapter 8).

Frequency domain pitch analysis techniques take advantage of the harmonic structure inherent in the (short-term) spectral representation of voiced speech. The pitch corresponds to the fundamental component of these frequency striations. Sometimes the fundamental frequency component is not present (for instance, in some sounds emitted by certain musical instruments (cf. Houtsma and Goldstein, 1972) and in speech signals that have been transmitted through a limited bandwidth telephone system). In other cases it is not clearly defined (perhaps because the first formant frequency is close to the fundamental frequency). The fundamental frequency in both the above situations can be determined by calculating the difference in frequency between adjacent peaks in the spectrum or by calculating the common factor of two or more widely spaced peaks (cf. Harris and Weiss, 1963). Other techniques include the use of the *cepstrum*, which is the Fourier transform of the logarithm of the power spectrum (§3.3.2). The "fine structure" (which is what the striations caused by the pitch are called) is periodic at a rate equal to the pitch period, and so produces a peak at that "quefrency" in the cepstrum. The pitch period for that segment of speech is determined by locating the position of the cepstral peak (Rabiner and Schafer, 1978, §7.3).

Pitch estimates obtained by any of the above techniques can often be improved by further processing. For example, a pitch estimate that is half or double the length of

the estimates of the surrounding pitch intervals is obviously wrong and should therefore be discarded (cf. Rabiner *et al.*, 1975; Sutherland *et al.*, 1988).

3.2 Linear prediction analysis

The source-filter model of speech production (§2.3.1.4) provides a simple framework for extracting features from speech signals. The characteristics of the source signal (loudness, voiced or unvoiced excitation, and pitch if voiced) can be estimated by straightforward means as described in §3.1. All that remains in order to complete a parametric description of the speech signal is to describe the parameters of the filter component. One successful method of parametrising the speech filter is by means of *linear predictive coding* (LPC). In this section, the mathematical basis of LPC analysis is briefly described (§3.2.1) and then (in §3.2.2) some of the details of implementing the analysis are discussed. The text by Markel and Gray (1976) provides extensive details of all aspects of the mathematics and implemention of LPC analysis.

3.2.1 Mathematical description

The source-filter model introduced in §2.3.1.4 expresses a speech signal as a convolution between an excitation, or source, component and a time-varying filter component. In order to justify the approximations used in the linear prediction (LP) model, it is useful to expand the source-filter equation introduced in §2.3.1.4 and describe the general characteristics of each term. In the z-domain, (2.1) can be expanded as (Markel and Gray, 1976, §1.3)

$$S(z) = E(z)G(z)V(z)L(z) \tag{3.5}$$

where $E(z)$ is the excitation, $G(z)$ is the glottal shaping filter, $V(z)$ is the vocal tract transfer function, and $L(z)$ describes the characteristics of sound radiation from the lips.

The excitation function $E(z)$ describes the energy source for the speech signal. It comprises a quasi-periodic impulse train for voiced sounds and white noise for unvoiced, both of which have an essentially flat spectrum. As described in §2.3.1.1, glottal pulses vary in shape from one to another. However, the spectral shape of an average glottal pulse can be approximated as a -12dB/octave slope (cf. Flanagan, 1972, p233; Sundberg and Gauffin, 1979). This is equivalent to the transfer function of a filter with two poles (Atal and Hanauer, 1971):

$$G(z) = \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})} \tag{3.6}$$

where g_1 and g_2 are close to unity. The effect of lip radiation can likewise be approximated by a first order filter with a +6db/octave spectral slope (Flanagan, 1972, §6.25):

$$L(z) = 1 - z^{-1} \tag{3.7}$$

The transfer function of a lossless acoustic tube (§2.3.1.3) contains only poles, so is described by

$$V(z) = \frac{1}{\sum_{i=1}^{K} a_i z^{-1}} \tag{3.8}$$

where K is the order of the filter.

By combining (3.6), (3.7), and (3.8), and noting that the zero in $L(z)$ is effectively cancelled by one of the poles in $G(z)$, a simplified all-pole model of the "speech shaping filter" can be derived:

$$G(z)V(z)L(z) = \frac{1}{A(z)} \tag{3.9}$$

where

$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-1} \tag{3.10}$$

is an all-pole filter which models the vocal tract shape, glottal pulse shape, and lip radiation characteristics. The all-pole speech model is obtained by substituting $A(z)$ into (3.5):

$$S(z) = E(z)\frac{1}{1 - \sum_{i=1}^{P} a_i z^{-1}} \tag{3.11}$$

By transforming (3.11) into the sampled time domain, the *speech prediction model*

$$s[n] = e[n] + \sum_{i=1}^{P} a_i s[n-i], \tag{3.12}$$

is obtained. This can also be expressed as

$$e[n] = s[n] - \hat{s}[n]. \tag{3.13}$$

where $\hat{s}[n]$ is the predicted value of the n^{th} speech sample.

In (3.12), the value of the n^{th} speech sample is predicted by a weighted sum of the previous P samples. The *prediction error* $e[n]$ (also called the *residue*) is seen to be equivalent to the excitation that is necessary to reconstruct $s[n]$ from the LPC coefficients.

Linear prediction entails choosing a set of coefficients $\{a_i\}$ such that the prediction error is minimised. Usually, a least-squares minimisation is performed over a number of sample values, although Denoel and Solvay (1985), for instance, employ a mean absolute error criteria, while Lee (1988) develops a coefficient estimation algorithm based on a general error weighting criterion. Least-squares minimisation results in the set

$$\sum_{i=1}^{P} a_i \mathbf{C}_{ik} = -\mathbf{C}_{0k} \qquad k = 1, 2 \ldots P \tag{3.14}$$

of linear equations, which can be straightforwardly solved using standard techniques (Markel and Gray, 1976). The matrix $\mathbf{C}$ is composed of short-term autocorrelation or covariance coefficients, defined by (Markel and Gray, 1976, pp13-15)

$$\mathbf{C}_{ij} = \sum_{n=n_0}^{n_1} s[n-i]s[n-j]. \tag{3.15}$$

In the *autocorrelation method*, n_0 and n_1 are effectively set to $-\infty$ and $+\infty$ respectively. However, the speech signal is windowed so that it is zero outside the interval $0 < n < N$. Hence (3.15) is equivalent to the short-term autocorrelation, with

$$C_{ij} = R_N[|i-j|], \tag{3.16}$$

where $R_N[k]$ is the autocorrelation of the speech segment containing N samples (Rabiner and Schafer, 1978, p402). Because of the truncation, each coefficient in $R_N[k]$ is

formed from a different number of summation terms. This introduces errors into the autocorrelation coefficients and hence into the predictor coefficients. This error can also be understood to occur because the predictor filter is effectively trying to match the sudden discontinuity at each end of the analysis segment (Rabiner and Schafer, 1978, §8.1.1). Hence, the prediction error $e[n]$ is large for the first and final P samples of the segment. It is thus good practice to window the segment with a tapered window (§1.3.1.1) in order to reduce the effects of the errors at the ends of the segment. Some of the consequences of this windowing are discussed in §3.2.2.

Another method of LPC analysis is the *covariance method.* In this approach, the summation limits in (3.15) are arranged so that each member of the set C_{ij} of coefficients is computed from the same number of summation terms. This can be accomplished by setting $n_0 = 0$ and $n_1 = N - 1$, and allowing the summation in (3.15) to include speech samples beyond the limits n_0 and n_1. Therefore, by contrast with the autocorrelation method, there are no discontinuities at the segment boundaries. Hence it is not necessary (or desirable) to window the signal (Rabiner and Schafer, 1978, §8.1.2). Various pertinent aspects of both these methods, including details of their implementation, are discussed further in §3.2.2.

A third method of obtaining the LPC coefficients is via a "lattice filter" formulation (Rabiner and Schafer, 1978, §8.3.3), which models the resonances in an acoustic tube model of the vocal tract. It therefore generates a set of coefficients different from the predictor coefficients obtained by the technique described above. The coefficients can be straightforwardly transformed from one formulation into the other, however. The details of the lattice filter method are presented by Friedlander (1982)

Linear prediction can also be developed as a spectral matching scheme, rather than a time domain predictive coder. In this approach, the parameters are chosen so that the model spectrum fits the speech spectrum according to a *maximum likelihood* criterion (Itakura and Saito, 1968; Schroeder, 1984). Such a model leads to the same set of equations as derived above. The details are beyond the scope of this thesis, and so interested readers are referred to Markel and Gray (1976, §2.2) or Schroeder (1984) for a more complete treatment.

3.2.2 Implementation techniques and considerations

In the autocorrelation method, the matrix $\mathbf{C}$, introduced in §3.2.1, is Toeplitz (symmetric with all elements along a given diagonal equal), which means that it can be solved using the Levinson-Durbin recursive algorithm (Rabiner and Schafer, 1978, §8.3.1). In the covariance method, the matrix $\mathbf{C}$ is symmetric and so Cholesky decomposition can be employed to solve (3.14) (Rabiner and Schafer, 1978, §8.3.2). Optimised computer programs for implementing both of these algorithms are presented by Markel and Gray (1976, §9.3).

As mentioned in §3.2.1, the autocorrelation method of analysis usually requires that the signal be windowed by a tapering (e.g. a Hamming) window. The use of a window reduces the fluctuations in the analysis error E (and hence in the computed coefficients) that occur as the analysis frame is moved, sample by sample, along the speech signal (Rabiner *et al.*, 1977). Large analysis errors occur because of the effective discontinuities at the ends of the analysis frame, as described in §3.2.1. However, although windowing reduces both the value of E itself and its sensitivity to the position of the analysis frame relative to the excitation, the error between the LPC coefficients and the actual speech signal is considerably greater than if no windowing is employed (Brieseman, 19XX). This is because the LPC coefficients match the *windowed* speech signal, which is significantly different from the *actual*, unwindowed signal.

In the covariance method of analysis it is not usual to window the speech,

because the summation limits in (3.15) are arranged so that the end-effects, which occur in the autocorrelation method, are absent. However, the coefficients obtained via the covariance approach are strongly affected by the position of the analysis frame relative to the glottal excitation pulses (Rabiner *et al.*, 1977). This is because the assumption that $e[n]$ (in (3.12)) is negligible is violated during the portions of the pitch interval wherein the excitation occurs. Hence, the coefficients are adversely affected by the presence of such excitation in the analysis frame. In order to obtain reliable coefficients, which accurately model the resonances of the vocal tract, the analysis window should be aligned with the *closed glottis interval* (CGI). During this interval, no excitation occurs, and so the speech signal is due solely to the decaying resonances of the vocal tract (cf. Makhoul and Wolf, 1972). Methods of locating the CGI are discussed in §3.4.1 (also see §3.3.3 for more discussion of the importance of aligning the analysis frame correctly).

The main advantage of the autocorrelation method over the covariance method is that the estimated coefficients are guaranteed to produce a stable all-pole filter (Rabiner and Schafer, 1978, §8.4). However, the covariance method produces coefficients that, when they are stable, more accurately represent the speech signal (Rabiner and Schafer, 1978, §8.5). Stability can usually be ensured by suitable selection of the analysis interval (as discussed in the previous paragraph), but this increases the computation required for the analysis. Hence the autocorrelation method is commonly employed in low data rate speech encoding schemes, where fast computation and stability of the LPC filter are required (Witten, 1982, p135). For applications where accuracy is a more important consideration, such as glottal inverse filtering (§3.4.1) or vocal tract area estimation (§3.2.3), the covariance method, in conjunction with a technique for estimating the correct analysis frame, is often employed (Markel and Gray, 1976, §4.4.1).

In order to adequately model speech sounds, the order P of the speech filter $A(z)$ must be sufficient to account for the vocal tract resonances, the glottal pulse shape and lip radiation. In addition, if the vocal tract filter contains zeros (such as for nasal or fricative sounds), additional terms are required in $A(z)$ to approximate their effect. Because the LPC coefficients can be related directly to the geometry of the vocal tract (see §3.2.3), enough coefficients must be computed to specify the entire length of the tract. Sound waves require about 1ms to propagate the 17cm average length of the vocal tract for adult male speakers. Hence the duration encompassed by the coefficients should also be of this order (Wakita, 1973). Practical values for the number of coefficients are therefore suggested as "the sampling rate in kHz plus 4 or 5" (Markel and Gray, 1976, p154). The 4 or 5 additional coefficients are required to adequately represent the departures from the all-pole model that are mentioned above (cf. Makhoul and Wolf, 1972).

The speech signal is usually differentiated before LPC analysis is performed upon it in order to improve the numerical stability of the analysis (Gray and Markel, 1974). This also effectively cancels the combined effects of the glottal shaping filter and lip radiation (§3.2.1). The speech filter $A(z)$ then effectively models the vocal tract only. The assumptions of negligible average excitation that lead to (3.14) are thus more closely met than if undifferentiated speech is analysed. When speech is reconstructed from such coefficients, it must be integrated (low-pass filtered with a filter having a -6dB/octave slope) in order to restore its correct spectral balance. If the intention of LPC analysis is to extract the shape of the vocal tract from the speech (§3.2.3), a first order differentiation is useful because the coefficients then match the vocal tract filter $V(z)$ more accurately (Wakita, 1973; Markel and Gray, 1976, §4.4)

The predictor coefficients are sensitive to small errors (such as those introduced by quantisation) because the roots of the predictor polynomial $A(z)$ are the poles of

the LPC speech filter. The filter is only stable if all its poles are inside the unit circle (§1.2.5.2). Because small differences in the values of the polynomial coefficients can cause large changes in the root positions, quantisation errors in the predictor coefficients can cause the LPC filter to become unstable. In order to guard against this source of instability, the coefficients are usually transformed into a different formulation, that is less sensitive to quantisation errors, before quantising them (Makhoul, 1975). §3.2.3 describes the different sets of coefficients that are commonly employed.

3.2.3 Alternative sets of LPC coefficients

The coefficients obtained by means of (3.14) are termed *prediction coefficients*, because they allow one to directly predict the value of a speech sample from knowledge of previous samples. By suitable transformations, other (equivalent) sets of coefficients can be obtained. In this section I introduce the coefficient sets that are commonly invoked, together with the transformations necessary to obtain them from the prediction coefficients. Further details of these coefficient sets can be found in the references cited.

The pole positions of the speech filter $1/A(z)$ can be evaluated by factoring the polynomial (3.10) to give

$$A(z) = \prod_{p=1}^{P}(1 - z_p z^{-1}) \tag{3.17}$$

where $\{z_p\}$ is the set of *LPC poles*. The speech filter $1/A(z)$ is stable providing that all the poles are inside the unit circle. The poles are the z-transform representation of the resonances of the vocal tract. The frequency $\hat{F}$ and bandwidth $\hat{B}$ corresponding to a pole at $z = z_p$ are given by

$$\hat{F} = \frac{f_s}{2\pi}\tan^{-1}\{\mathcal{I}\{z_p\}/\mathcal{R}\{z_p\}\}\text{Hz} \tag{3.18}$$

and

$$\hat{B} = -\frac{f_s}{\pi}\ln|z_p|Hz \tag{3.19}$$

respectively, where f_s is the sampling frequency (Atal and Hanauer, 1971).

As stated in §1.2.4, the DFT corresponding to the z-transform of a signal can be obtained by evaluating the z-transform function at equally spaced points around the unit circle. Hence the spectrum $|S(f)|^2$ of a speech segment can be obtained from the LPC filter $1/A(z)$ of that segment by evaluating

$$|S(f)|^2 = \frac{1}{|1 - \sum_{j=1}^{P} a_j e^{-2\pi ijf/f_s}|^2} \tag{3.20}$$

where f_s is the sampling frequency of the speech signal (Rabiner and Schafer, 1978, §8.6). Examples of speech spectra obtained by this method are presented in §3.3.3.

The cepstral coefficients $c[n]$ (§3.3.2) of the LPC all-pole filter can be obtained directly from the prediction coefficients $\{a_i\}$ by the recursive procedure

$$c[n] = a_n + \sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)c[n-k]a_k \quad n > 0 \tag{3.21}$$

where $a_0 = 1$ and $c[1] = a[1]$ (Atal, 1974).

The impulse response of the LPC filter can be straightforwardly obtained by evaluating (3.12) with $e[n] = \delta[n]$, where $\delta[n]$ is a unit impulse at $n = 0$.

The all-pole model of the vocal tract filter is equivalent to a lossless resonating tube (§2.3.1.3). Two types of coefficients which relate to the acoustic tube model can be derived from the LPC coefficients. If the acoustic tube is modelled as several concatenated uniform tubes, then the *area coefficients* characterise the cross-sectional area of each of the tubes (Wakita, 1973). In addition, the *reflection coefficients* describe the reflections that occur between each segment (Markel and Gray, 1976, §4.2). Reflection coefficients are equivalent to the *PARCOR* partial correlation coefficients that are produced as a byproduct of the autocorrelation method (Wakita, 1973).

The reflection coefficients k_j are related to the areas $\mathcal{A}_j$ of each section of the vocal tract by the relation (Markel and Gray, 1976, §4.2)

$$\mathcal{A}_j = \mathcal{A}_{j-1}\frac{1+k_j}{1-k_j} \tag{3.22}$$

with

$$\mathcal{A}_1 = 1 \tag{3.23}$$

The reflection coefficients can be obtained from the prediction coefficients by the following "backward" recursion, where $i = P \ldots 1$ (Rabiner and Schafer, 1978, §8.8):

$$\begin{aligned} k_i &= \alpha_i^{(i)} \\ \alpha_j^{(i-1)} &= \frac{\alpha_j^{(i)} + \alpha_i^{(i)}\alpha_{i-j}^{(i)}}{1-k_i^2} \quad 1 \le j \le i-1 \end{aligned} \tag{3.24}$$

The coefficients $\{\alpha_j^{(P)}\}$ are set equal to the predictor coefficients $\{a_j\}$:

$$\alpha_j^{(P)} = a_j,\ 1 \le j \le P. \tag{3.25}$$

A "forward" recursive procedure is invoked to obtain the predictor coefficients from the reflection coefficients. Thus, for $i = 1 \ldots P$,

$$\begin{aligned} \alpha_i^{(i)} &= k_i \\ \alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i\alpha_{i-j}^{(i-1)} \quad 1 \le j \le i-1, \end{aligned} \tag{3.26}$$

with the predictor coefficients given by (3.25) once again.

If the area coefficients that are obtained from the speech signal via the LPC method are to accurately match the true vocal tract area, it is important that the effects of lip radiation and glottal excitation are separately accounted for. This can be accomplished by applying a simple +6dB/octave pre-emphasis to the speech signal (Wakita, 1973), although, as discussed by Sondhi (1979), the sensitivity of the computed areas to the particular compensation employed means that techniques such as CGI LPC analysis should be invoked to improve the estimation of the coefficients (cf. §3.3).

3.3 Frequency domain analysis

As described in §2.2.2 and §2.3.2, the cochlear extracts spectral information from sounds entering it. Furthermore, speech is produced by a mechanism (the vocal tract) which is usefully modelled as a number of resonant chambers (§2.3.1.3). Hence many approaches to analysing speech signals are based on extracting information from a spectral representation of speech. In §3.3.1 I discuss the concept of a time-varying spectrum, which can be employed to represent the spectral content of a speech signal. §3.3.2 describes the cepstrum, which is useful for revealing some of the structure inherent but not readily apparent in the spectrum, and §3.3.3 introduces some of the ways that spectral analysis techniques are employed to extract information from a speech signal.

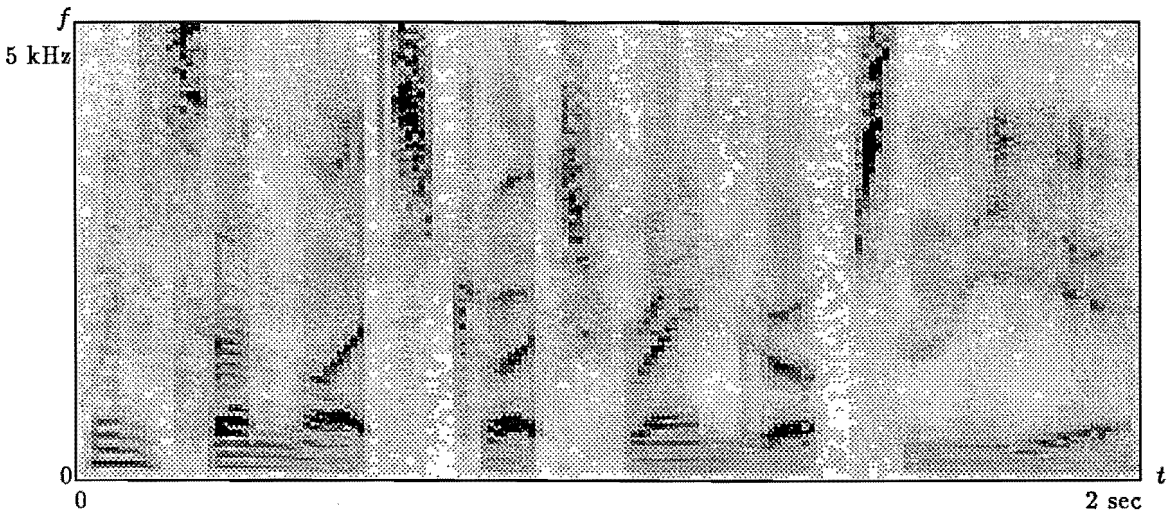


Figure 3.2. Time-varying spectrum (spectrogram) of the first two seconds of the utterance AM-RAIN1. A 256 sample long Hamming window was used to delineate each segment. The image density represents energy, with white being lowest and black highest.

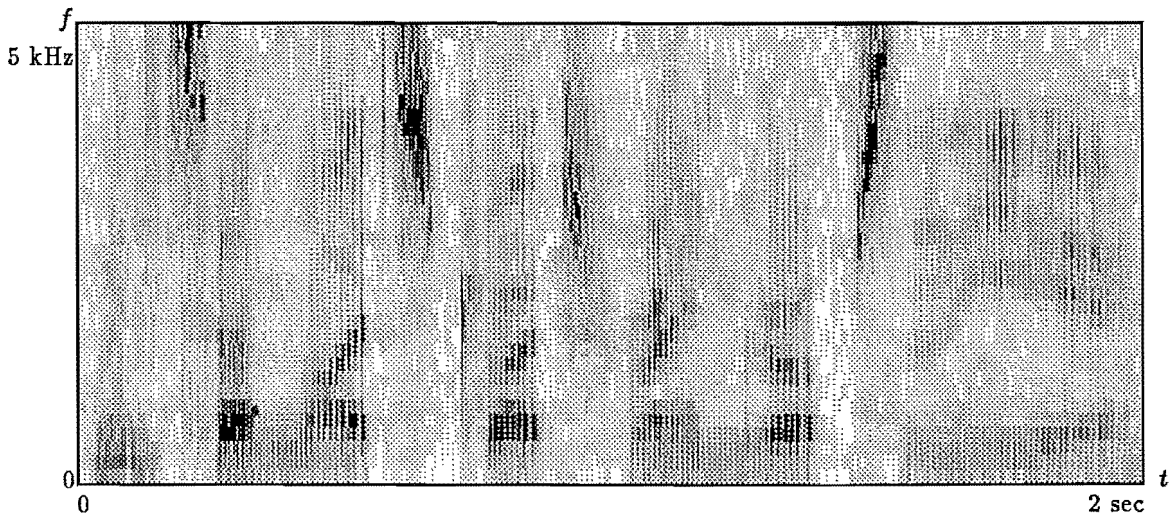


Figure 3.3. Time-varying spectrum (spectrogram) of the first two seconds of the utterance AM-RAIN1. A 64 sample long Hamming window was used to delineate each segment. The image density represents energy, with white being lowest and black highest.

3.3.1 Spectral representation of speech

Speech can be considered as a signal whose (short-term) spectral content is continually changing (due to the movements of the articulators). However, for segments of short duration (of the order of 10–20ms), the speech signal is essentially the output of a time-invariant system (§2.3.1.4). It can therefore be considered to be *invariant* over such short intervals (cf. §1.2.5.4). The *time-varying* spectrum (or *spectrogram*) is constructed from the power spectra of successive segments. It can be visualised as a two-dimensional signal, with frequency along one dimension and time along the

other. Note that this "frequency" is different from "conventional frequency", as defined in §1.2.1, because it is defined as the Fourier transform of only a short segment of the signal. Hence the frequency axis of the time-varying spectrum is properly referred to as the "short-term frequency" axis (see §1.2.1). Fig.3.2 shows an example of a spectrogram calculated from the utterance AM-RAIN1. The horizontal and vertical axes denote time and frequency respectively, while the image density at each point in the spectrogram indicates the portion of the speech energy at that point in time and frequency. In the remainder of this section, I describe some of the details of spectrographic speech analysis. Further details are presented by, among others, Rabiner and Schafer (1978, Chapter 8) and Tribolet and Crochiere (1979).

The time-varying spectrum of a (discrete) signal $s[m]$ is defined by (Flanagan, 1980)

$$S_n[k] = \sum_{m=-\infty}^{\infty} h[n-m]s[m]e^{-2\pi ikm} \tag{3.27}$$

where $S_n[k]$ is the (discrete) short-term spectrum at the instant n and $h[m]$ defines the *window* (see §1.3.1.1) over which the transform is obtained. The time-varying spectrum $S_n[k]$ can be interpreted either as the output of a bank of filters, or as the Fourier transform of the segment of speech $h[n-m]s[m]$ (Tribolet and Crochiere, 1979). In the *filter-bank* interpretation, (3.27) is viewed as a convolution between a low-pass filter response $h[n]$ and the signal $s[n]$ modulated by a sinusoid of frequency k:

$$s_k[n] = h[n] \odot [s[n]e^{-2\pi ikm}] \tag{3.28}$$

where $s_k[n]$ is the output of the filter with centre-frequency k. In the *block transform* interpretation, $S_n[k]$ is viewed as the Fourier transform of the segment of speech delineated by the extent and position n of the window $h[n-m]$. If a tapered window (§1.3.1.1) is employed to reduce the effects of leakage, it is necessary to overlap adjacent segments. This is because such a window reduces the contribution from speech samples near to the ends of the segment. Hence the segments need to be spaced closer than the actual duration of the window. The next paragraph discusses these "overlap requirements" further, but from the viewpoint of the filter-bank interpretation of the time-varying spectrogram.

Because the filter-bank interpretation of the time-varying spectrum indicates that $h[n]$ performs a filtering function as well as a windowing function, it is necessary (for accurate determination of the short-term spectra) that the frequency response of $h[n]$ be well behaved. In particular, it should not have high side-lobes that may cause leakage of strong spectral components into other than the correct filter band. In addition, if the individual sub-bands are sampled at a lower rate than the original signal is, the magnitude of the Fourier transform of $h[n]$ should be small enough, at frequencies greater than half the new sampling frequency, that aliasing does not occur. The approximate bandwidth of a Hamming window (§1.3.1.1), of length L samples, is (Rabiner and Schafer, 1978, p264)

$$B_H = \frac{2F_s}{L}, \tag{3.29}$$

where F_s is the frequency at which the speech signal is sampled.

Fig.3.3 shows a spectrogram, of the same utterance as that to which Fig.3.2 relates, but computed with a wide instead of a narrow bandwidth filter ("short" instead of "long" window). Although the resolution with which true spectral features are revealed is very much reduced in Fig.3.3, rapid *transitional* features are revealed much more clearly. §3.3.3 describes in more detail some of the methods of abstracting useful information from "spectrographic" representations of speech signals (see also Koenig *et al.*, 1946)

3.3.2 The cepstrum

A convolution in the time domain between two signals is represented in the frequency domain by a multiplication between their respective spectra (§1.2.5.1). The logarithm of the composite spectrum is thus composed of a sum between the two components. Hence the components can be separated by linear filtering if their individual *cepstra* do not overlap. The complex cepstrum $c(\tau)$ of a signal $s(t)$ is defined by (Bogert *et al.*, 1963)

$$c(\tau) = \mathcal{F}^{-1}\{\log(S(f))\} \tag{3.30}$$

where $S(f)$ is the spectrum of $s(t)$ and the independent variable τ is often termed *quefrency* to distinguish it from the time variable, t. Because $S(f)$ is complex, the logarithm is defined as

$$\log(S(f)) = \log(|S(f)|) + i\phi(f) \tag{3.31}$$

where $\phi(f)$ is the phase of $S(f)$. The complex cepstra is only required if the spectral phase is important (for example, if the time domain signal is to be reconstructed). If not, the cepstrum can be calculated from the magnitude spectrum, in which case (3.30) and (3.31) are replaced by their real equivalents (Oppenheim and Schafer, 1975, Chapter 10).

The cepstral components are related to the log spectrum of a signal in the same way that the spectral components are related to the time domain signal (hence the terms "cepstrum" and "quefrency" that were introduced by Bogert *et al.*, 1963). The low quefrency components therefore characterise the smoothly varying components of the spectrum, while the high-quefrency components characterise closely spaced (harmonic) components of the spectrum. The (complex) cepstrum has been employed for blind deconvolution of signals whose constituent components differ in this manner (Oppenheim *et al.*, 1968). One such signal is of course speech, because the pitch periodicities introduce closely spaced harmonics into the spectrum, while the formants have a much smoother spectral shape (§3.3.3). Hence these two components can often be separated in the cepstral domain (Rabiner and Schafer, 1978, §7.2). Fig.3.4 shows time domain, spectral domain and cepstral domain representations of a typical segment of voiced speech. Cepstral analysis of speech is invoked for pitch estimation (§3.1.3), extracting smoothed estimates of the formant frequencies (§3.3.3), low data rate speech encoding (§3.5), and speech recognition (§3.6.1). Further details of the application of cepstral analysis techniques to speech signals are covered by Rabiner and Schafer (1978, Chapter 7).

Cepstral analysis is often referred to as a type of *homomorphic signal processing* (cf. Oppenheim *et al.*, 1968). A homomorphic system is one that obeys a generalised superposition principle (cf. §1.1.3). It can therefore described by a "generalised convolution", similar to that introduced in §1.1.3 for linear systems. By suitable transformations the convolution can be represented as a summation of the two components. Cepstral analysis is one technique for performing this transformation. The details of generalised homomorphic signal analysis are beyond the scope of this thesis, but they are treated by Oppenheim and Schafer (1975, Chapter 10).

3.3.3 Extracting information from a speech spectrum

The time-varying spectrum (whether calculated by Fourier or LPC analysis) of a speech signal conveys information about the types of sounds being uttered (§2.1.4). The main features of interest in the time-varying spectrum are the *formants*, or spectral peaks, which correspond to the resonances of the vocal tract (§2.3.1.3). As indicated in §2.1.4, the frequencies of the first two formants, or vocal tract resonances, are sufficient to

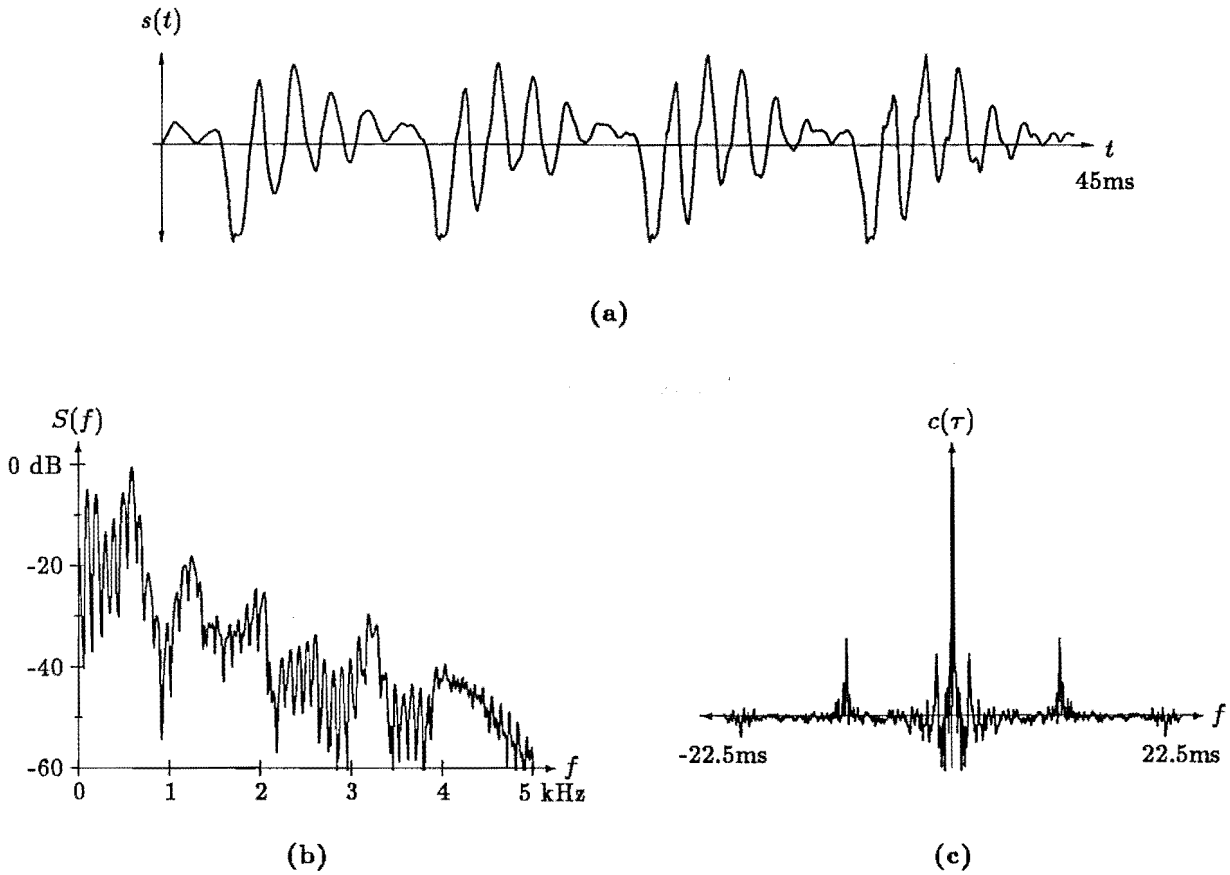


Figure 3.4. Example of cepstral analysis of a short segment of voiced speech. **a:** Time domain, **b:** Spectral domain and **c:** Cepstral domain representations of the same speech segment (from utterance AM-RAIN1).

identify each of the vowels. The consonants are identified by (among other features) the formant "transitions" (cf. §2.1.4.2). Hence it is necessary to resolve both the frequencies of the formants and the timing of the transitions in order to reliably extract phonetic features from the time-varying spectrum of a speech signal. There is an unavoidable trade-off between the spectral and the transitional resolution that can be obtained (see §1.2.5.4, §1.3.1.1, and Gabor, 1946). However, the human cochlear also performs a "spectrographic" type of analysis in order to detect sounds (§2.2.2.1), and so it is subject to the same "time-frequency" trade-off. It is therefore only necessary to ensure that the spectrographic analysis employed in the feature extraction process has similar characteristics to that employed by the human cochlear.

In order to make the spectrographic analysis of speech mimic the human auditory system more closely, the "raw" spectrogram can be modified (i.e. *perceptually weighted*) according to the characteristics of human hearing (see §2.2.2). For example, Cohen (1989) weights the short-term spectra of the speech sounds by the auditory characteristics of loudness perception, critical bands, and neural adaptation (cf. §2.2.2.2). Ghitza (1987) describes a system in which the short-term spectrum is divided into 100 overlapping sub-bands. The sub-bands are shaped according to the characteristics of the human peripheral auditory system (see §2.2.2.2). Each sub-band is considerably wider than the reciprocal of the effective duration of the analysis interval (§3.3.1), allowing many individual frequency components to be identified within each sub-band. The dominant frequency component in each sub-band is then identified, with those fre-

quency components appearing in several adjacent bands being selected as characteristic features of the speech sound. This selection of the spectral peaks that occur in several "auditory bands" is supposed to model the patterns of nerve firings that occur in the auditory nerve (cf. Allen, 1985; Greenberg, 1988).

The formant frequencies can be computed directly from the LPC coefficients, by invoking (3.18). This gives the frequencies of the "poles" of the LPC speech filter, which correspond to resonances in the vocal tract. However, the LPC filter generally contains more poles than there are formants, because additional poles are required to adequately model a speech signal (§3.2.2). In addition, for segments of speech that are not all-pole (such as nasal or fricative sounds), it is not clear what the pole frequencies represent *vis a vis* the vocal tract resonances (McCandless, 1974; Rabiner and Schafer, 1978, p450). Fig.3.5*a* shows the pole frequencies corresponding to the LPC coefficients of the utterance AM-RAIN1. The formant frequencies can be seen, but they are obscured by the many other poles of different frequencies. For these reasons determining the "true" formant frequencies from the pole positions is not a trivial task. Techniques for "tracking" the formants from frame to frame are necessary to determine which poles correspond to the formants and which to other features of the speech signal. McCandless (1974) and Markel and Gray (1976, Chapter 7) present further details on these techniques.

The formant frequencies can also be obtained by computing the DFT from the LPC coefficients, as defined by (3.20). The formants correspond to the spectral peaks of the DFT. Fig.3.5*b* shows the spectrogram computed in this manner from the utterance AM-RAIN1. In some cases the exact frequencies of the formants can be difficult to determine because the spectral peaks are too wide. They can be made more pronounced by evaluating the LPC filter around a circle in the z-plane, centred on the origin, but with a radius less than unity. Because such a circle is closer to the poles than the unit circle (§3.2.3), the peaks in the computed spectrum are rendered much sharper (McCandless, 1974; Duncan and Jack, 1988).

Another obstacle in the path of reliable formant estimation is the influence of the glottal excitation. As described in §2.2.1 and §2.3.1, the speech signal is effectively composed of two components — the excitation and the vocal tract filter. For voiced sounds, the excitation consists of a quasi-periodic train of pulses. The estimated formant frequencies are influenced both by the periodicity of this excitation, and by the spectral content of the pulses (cf. Makhoul, 1973).

Fig.3.6*a* shows the spectrum obtained from a segment of speech of 45ms duration (which is approximately equal to four pitch periods). As shown, the positions of the formant peaks are obscured because of the pronounced "ripple" in the spectrum. This arises from the strongly harmonic nature of voiced speech. One method of removing this ripple is to deconvolve the pitch component out by *cepstral smoothing* (§3.3.2). Fig.3.6*b* shows the result of smoothing the spectrum in Fig.3.6*a* by discarding all but the first 4ms of the cepstrum. Another method of removing the "pitch ripple" is to use LPC analysis since, as implied in §3.2.1, this effectively ignores the pitch periodicities because they do not conform to the all-pole model (Makhoul, 1973). Fig.3.6*c* and *d* show the spectra obtained in this manner when 10 and 20 coefficients respectively are employed in the LPC analysis.

The spectral content of the glottal excitation influences the formant estimates by altering the frequencies of the observed peaks in the spectrum. This is especially severe for LPC analysis when a short analysis frame is employed (Markel and Gray, 1976, §7.6.3). The estimated formant frequencies then vary according to the position of the analysis frame relative to each pitch interval. The formant frequencies can be made more consistent by aligning each analysis frame with each glottal excitation pulse by

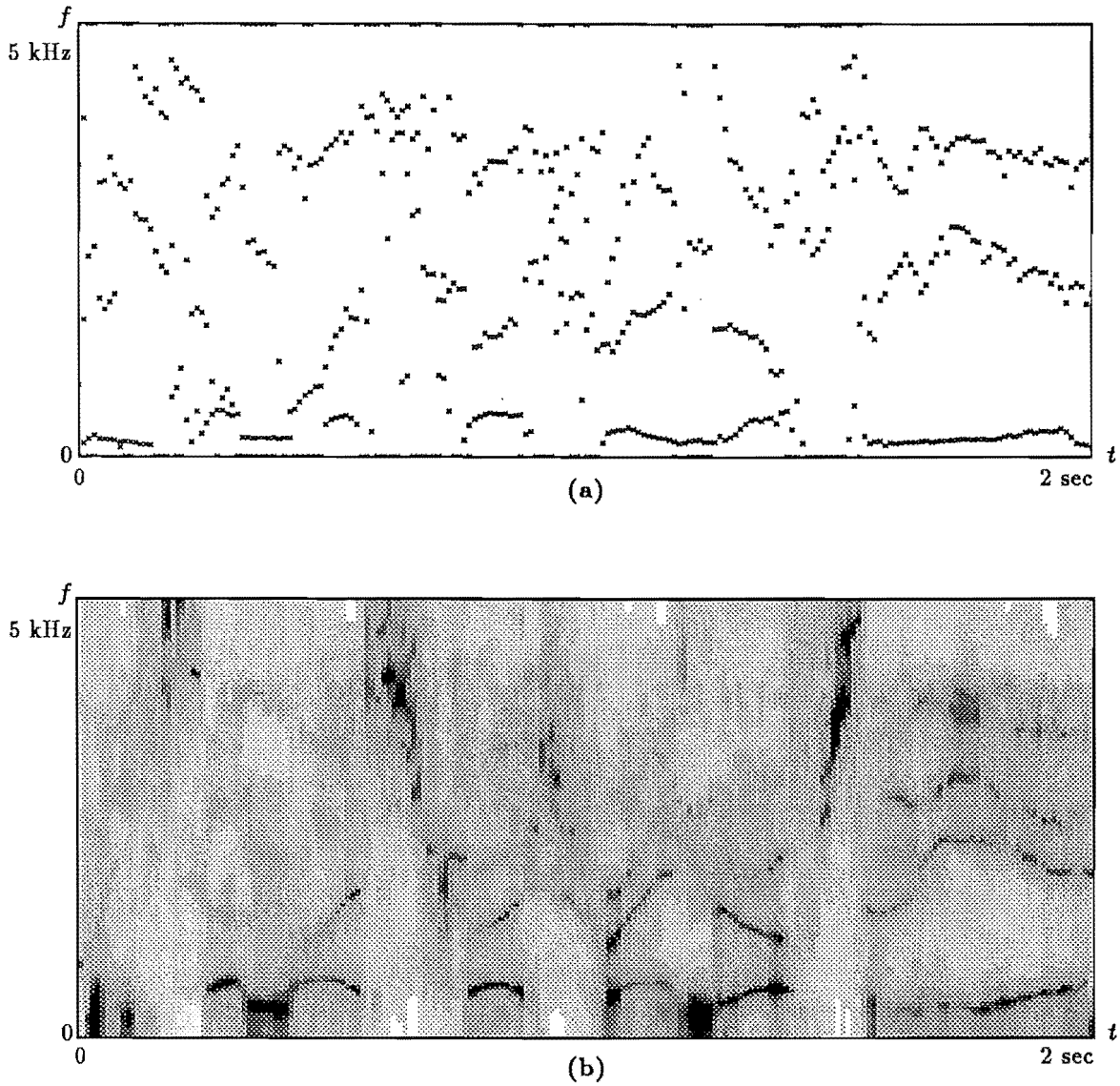


Figure 3.5. Spectrographic analysis via linear prediction, of the first two seconds of the utterance AM-RAIN1. A segment size of 128 samples, Hamming window, and the autocorrelation method of LPC were all employed to compute the coefficients. **a:** Pole frequencies evaluated by means of (3.18). Six LPC coefficients were employed. **b:** The spectrogram, computed by evaluating the LPC filter $1/A(z)$ at 128 points equally spaced around the upper half of the unit circle (see (3.20) in §3.2.3). Sixteen LPC coefficients were employed for each spectrum.

means of a suitable pitch algorithm (termed *pitch synchronous* analysis). Paliwal and Rao (1981) present such a method, whereby the pitch periodicity is taken explicitly into account by setting the duration of each analysis frame equal to the pitch period. Each analysis frame is positioned synchronously with the start of each pitch interval. Their technique provides better estimates of the spectral content of synthetic speech than do conventional autocorrelation or covariance methods.

"Pitch synchronous" analysis helps to make the formant estimates more consistent, which is especially important for speech recognition purposes. However, applications such as glottal waveform extraction (see §3.4), or estimation of the vocal tract shape (cf. Sondhi, 1979) require that the formant estimates represent the resonances

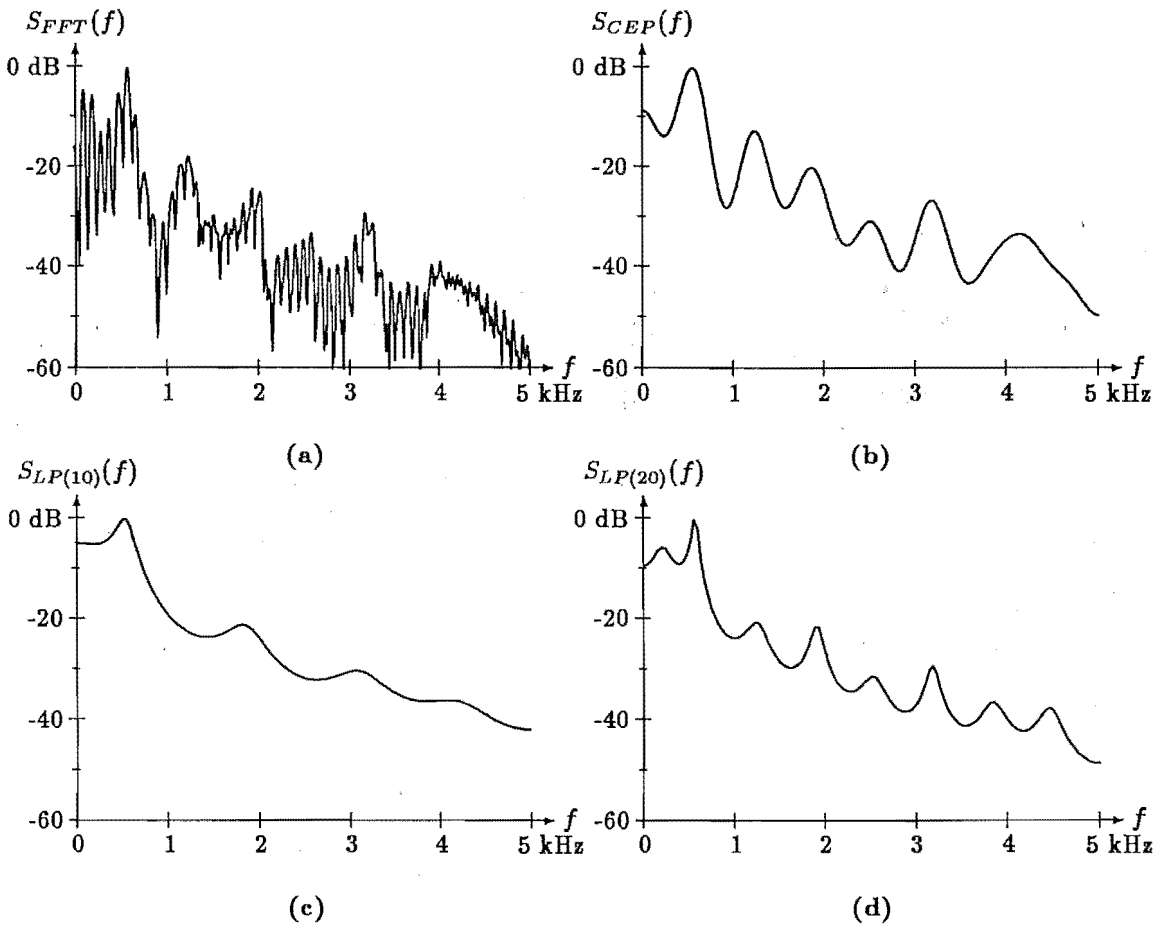


Figure 3.6. Smoothed spectrum of segment of speech shown in Fig.3.4*a*. The spectrum was smoothed by: *b* Computing the (real) cepstrum, discarding the coefficients with quefrency greater than 4ms, and transforming back to the Fourier domain; *c* and *d* Extracting (10 and 20 respectively) LPC coefficients from the speech signal then converting them to spectra by evaluating (3.20).

of the vocal tract alone. In order to be assured of this, the analysis window should be positioned in the *closed glottis interval* (CGI) of the excitation waveform (Wood and Pearce, 1989), which is when the glottis is (assumed to be) closed (see §3.4.1). The formant frequencies can be estimated more accurately when the glottis is closed because the effects of the sub-laryngeal airways on the vocal tract resonances are absent (cf. Makhoul and Wolf, 1972). The open glottis changes the resonant characteristics of the vocal tract, by introducing "zeros" that correspond to the resonances of the sub-laryngeal tract. The presence of these zeros leads to erroneous estimates of the vocal tract formant frequencies and bandwidths (cf. Makhoul and Wolf, 1972).

Pitch synchronous analysis can also be combined with Fourier methods of spectral estimation (Rabiner and Schafer, 1978, §6.6.1), in order to equalise, as far as possible, the effect of the glottal excitation in each analysis frame.

3.4 Glottal waveform estimation techniques

Voiced speech can be regarded as a convolution between a *glottal waveform* and a time-varying vocal tract filter (§2.3.1.4). Techniques for estimating the exact form of the glottal waveform from observations of the (voiced) speech signal are useful for several

reasons. Firstly, the quality of synthetic speech (§3.5.2) appears to be influenced by the type of excitation employed (Rosenberg, 1971; Holmes, 1973). By employing an excitation signal that is more closely related to the actual glottal waveform than the simple signals usually employed, more natural sounding speech can be synthesised (Holmes, 1973). A second application is the diagnosis of certain laryngeal disorders §3.6.4.1. Extracting the glottal waveform is a non-invasive method of assessing abnormalities of the vocal cords (Childers, 1990).

There are two main approaches to the problem of estimating the glottal waveform from the speech signal. Both approaches are effectively species of blind deconvolution. In the first approach, the vocal tract filter response is estimated, by taking into account knowledge of its general characteristics (see §3.2.1 and §3.4.1). The glottal waveform is then obtained by filtering the speech signal with the inverse of the vocal tract filter. Techniques that rely on this approach are discussed in §3.4.1. The second approach, discussed in §3.4.2, is to estimate the glottal waveform directly from the speech waveform.

An indication of the glottal waveform can also be obtained by tactile sensors which measure mechanical manifestations of the glottal vibration. Several of these types of techniques are mentioned in §3.4.1.

3.4.1 Inverse filtering techniques

The process of estimating the glottal waveform by inverse filtering of the speech signal proceeds by, first, estimating the vocal tract impulse response and, second, deconvolving that from the speech signal. The inverse filter is formed from the estimate of the vocal tract filter produced by either spectral or LPC analysis of the speech waveform. The accuracy of such an estimate of the glottal waveform depends upon the accuracy with which the estimated LPC filter models the vocal tract rather than the glottal excitation characteristics. Because of the effect of the lip radiation, which is approximately a first order differentiation, the waveform obtained by inverse filtering is usually the differential of the glottal waveform. Hence, some methods suffer from low frequency distortion caused by the requisite integration involved in obtaining the glottal flow waveform (cf. Veeneman and Bement, 1985). This means that the particular value of the computed waveform that corresponds to zero airflow in the actual waveform cannot be established unambiguously. Rothenberg (1973) overcomes this difficulty by measuring the air *flow* at the mouth instead of the sound *pressure*. In Rothenberg's technique, the air flow is measured by means of a special mask placed over the speaker's mouth. Several recent studies (cf. Karlsson, 1986; Fritzell *et al.*, 1986) have employed this same technique to investigate to what extent the glottis actually does completely close. Their results indicate that the amount of glottal closure varies between different speakers (see §3.6.4.1).

Early studies of inverse filtering characterise the vocal tract filter by an estimate of the formant frequencies and bandwidths obtained from a spectrogram of the speech signal (Miller, 1959). However, some manual adjustment of the inverse filter's parameters is required to obtain a "correct" result (Miller, 1959). The parameters are adjusted until the glottal waveform exhibits a "closed phase", a feature that is often observed in high-speed motion pictures of glottal vibrations (Fletcher, 1953). Rosenberg (1971) obtains the inverse filter by an (automatic) iterative procedure of varying the pole positions of a trial filter until the best match with the real spectrum is found. An initial glottal waveform is approximated by means of spectral zeros. A similar approach is taken by Oppenheim and Schafer (1968), who fit a series of resonators to the smoothed log-spectrum of the speech signal. The difference between the model log-spectrum and the actual log-spectrum of the speech becomes the log-spectrum of

the glottal waveform.

Most recent approaches to inverse filtering (cf. Veeneman and Bement, 1985; Wong *et al.*, 1979) have employed LPC estimates of the vocal filter. This is because, as emphasised in §3.3.3, LPC analysis employs an all-pole model that matches the formant resonances of the vocal tract. However, in order to ensure that the estimate of the vocal tract filter is not affected by the glottal excitation characteristics, it is advisable to perform the analysis only on the closed glottis interval of each pitch interval (see §3.2.2 and §3.3.3). While the requirement to identify the CGI before performing the analysis that reveals the glottal waveform seems somewhat circular, there are several techniques whereby the CGI can be approximately (and adequately) located beforehand. For instance, LPC analysis (using the covariance method) can be performed on analysis frames positioned at successive samples throughout the speech signal. The analysis error exhibits a periodic variation, according to the amount of excitation encompassed by the analysis frame (Rabiner *et al.*, 1977). So, the instant of glottal closure can be identified by the sharp fall in the analysis error that occurs when the analysis frame is positioned one sample after that instant in the speech signal (Wong *et al.*, 1979). The error then rises again as the analysis frame begins to overlap the next excitation pulse. The interval between the sharp fall and the subsequent rise is identified as the CGI.

Ananthapadmanabha and Yegnanarayana (1979) employ a similar technique to the above, except that they filter the LPC residue to identify the "epochs" corresponding to the locations of glottal closure. Because the glottal waveform has effectively been differentiated twice (by the effect of lip radiation and the differentiation performed as part of LPC analysis), any slope discontinuities in the flow waveform effectively become impulses in the LPC residue. Slope discontinuities occur at the instants of glottal closure because of the sudden cessation of flow. Even though other slope discontinuities may occur during the glottal flow, Ananthapadmanabha and Yegnanarayana (1979) find that they are still able to identify the instant of glottal closure for a wide variety of voices. However, for voices that are very noisy (see §3.6.4.1), the glottal closure epoch is not so easily identified.

Typical glottal waveforms obtained by inverse filtering segments of utterances from three different speakers are shown in Fig.3.7. These examples were generated using the approach of Wong *et al.* (1979), with the LPC analysis over the CGI performed by the covariance method. In each of Fig.3.7*a* and *b*, (i) shows the recorded speech signal, while (ii) shows the variation of LPC error signal as the analysis frame was moved along the (differentiated) speech signal. The "×" marked on the LPC error signal indicates the start of the analysis frame from which the inverse filter was computed. Each graph labelled with (iii) represents the differentiated glottal waveform, sometimes called the "equivalent glottal excitation" (Wong *et al.*, 1979). The glottal flow waveforms identified by (iv) are obtained by integrating the equivalent glottal excitation waveform.

Several recent studies (cf. Boves, 1984; Veeneman and Bement, 1985; Krishnamurthy and Childers, 1986) have employed additional information in the form of *electroglottographic* (EGG) signals in order to help identify the CGI. The use of signals additional to the speech signal is inconvenient for the application of speech analysis to low data rate speech encoding, but for clinical applications it is feasible. An EGG signal is obtained by placing electrodes on the skin, on either side of the voice-box, thus measuring the impedance across the glottis (Krishnamurthy and Childers, 1986). As the vocal cords vibrate, the impedance varies, and so the EGG gives an indication of the vocal cord contact area, which is related to the glottal area and glottal flow waveform (cf. Titze *et al.*, 1983). The EGG can also assist in accurate determination of the pitch frequency and VUV classification of the speech signal, because it indicates the actual vibration of the vocal cords, without any "distortion" applied to the signal

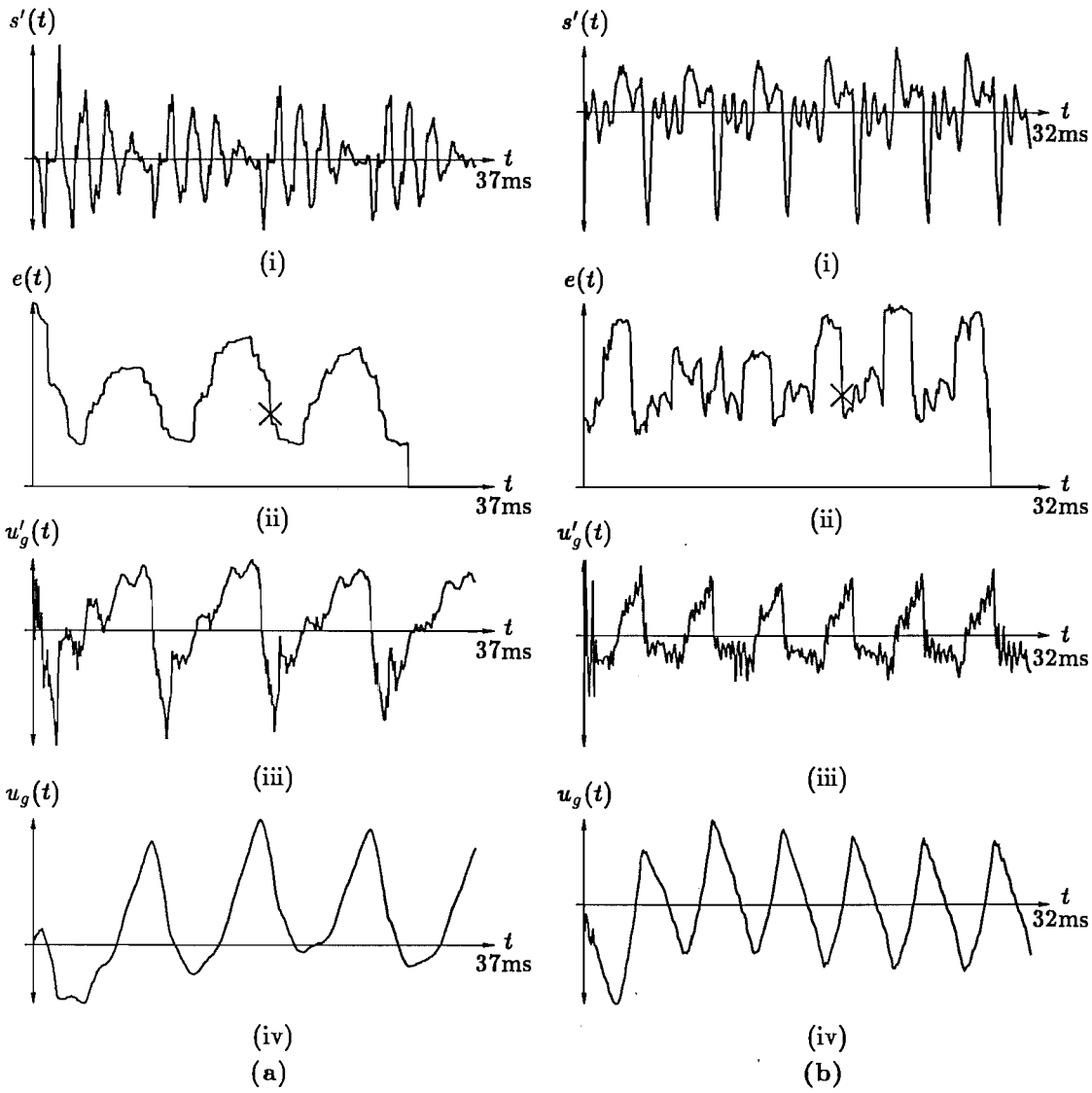


Figure 3.7. Typical glottal waveforms, obtained by inverse filtering of the speech signal. The segments of speech signal are from utterances **a:** AM-RAIN1 and **b:** TF-RAIN1. In each case the waveforms are: (i) Differentiated speech signal, (ii) LPC error, (iii) glottal excitation function, and (iv) glottal flow waveform. The LPC analysis frame is of length 5ms for the example in *a* and of length 2.5ms for that in *b*.

from the vocal tract resonances (Krishnamurthy and Childers, 1986).

Other tactile observations of the vocal cord vibration have involved the use of high-speed video and cinegraphy (cf. Fletcher, 1953; Anastaplo and Karnell, 1988; Childers, 1990), ultrasonic and photo-intensity measurements akin to EGG (Titze *et al.*, 1983; Boves, 1984), and the direct measurement of air pressure at various points in the vocal apparatus by means of miniature pressure transducers placed in the vocal tract (Boves, 1984). Such studies help to confirm the types of waveforms that are obtained by inverse filtering techniques and those that are predicted by models of vocal cord vibration (§2.3.1.1).

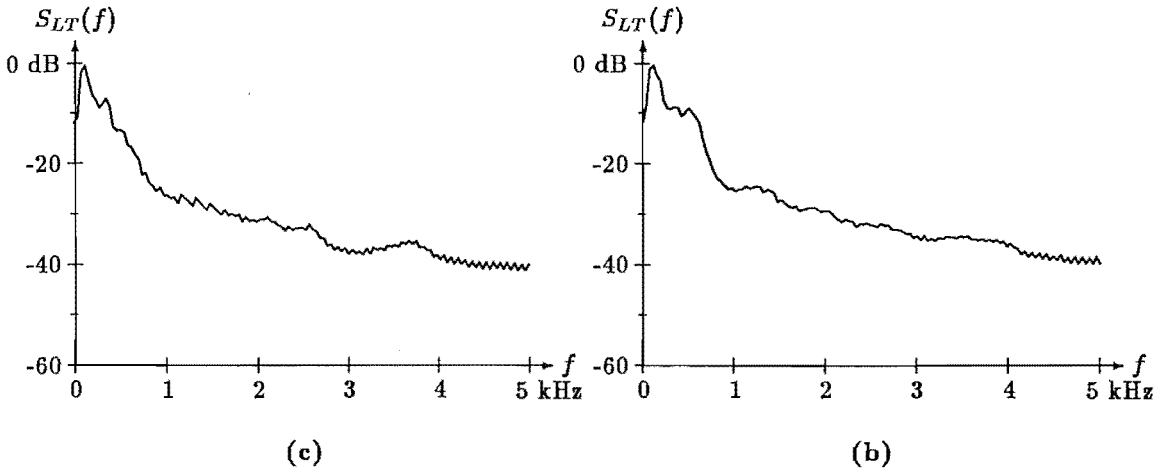


Figure 3.8. Long-term-average-spectra (LTAS) of utterances **a:** AM-WAL and **b:** TF-WAL.

3.4.2 Direct estimation

Only a few methods have been proposed for estimating the glottal waveform directly from the speech signal. Some of these techniques are "averaging" types of blind deconvolution, and therefore only estimate the average "glottal pulse". In this section I briefly describe two methods of "direct glottal estimation". The *shift-and-add* technique, which is also "direct", is neglected here because it is described in great detail in Chapter 4.

The *long-time average spectrum* (LTAS) is obtained by computing the average spectral content of the speech signal over a long (10–40seconds) period of time (Boves, 1984). Because the vocal tract filter varies continually and considerably during a typical utterance, while the glottal waveform spectrum is relatively invariant, the LTAS is effectively the spectrum of the average glottal excitation (Boves, 1984). Fig.3.8 shows the LTAS computed from utterances of two speakers. The LTAS of each utterance was obtained by dividing the voiced portions of the utterance into segments, each of duration 25.6ms. Each segment was windowed with a 3-term Blackman-Harris window and zero-extended (§1.3.1.1) to a total duration of 51.2ms before computing its power spectrum by means of the FFT algorithm.

The LTAS has been employed as a descriptor in a speaker recognition system (§3.6.2). It has also been investigated for use in a system for diagnosing laryngeal disorders (§3.6.4.1).

Another approach to estimating the glottal waveform, which is more akin to the straightforward inverse filtering methods described in §3.4.1, is to simultaneously extract the vocal tract parameters and the parameters of a glottal waveform model by some suitable optimisation technique. This is the approach used by Milenkovic (1986), who models the glottal waveform by parametrising its opening and closing transients. By incorporating the glottal waveform into the LPC analysis, both sets of parameters can be optimised by a single procedure.

3.5 Low data rate speech encoding techniques

Much of current speech research is aimed at investigating and developing techniques to encode speech signals in the most compact form possible, while still retaining the essential (in the context of a particular application) characteristics of the speech sounds (see §2.1.2, §2.1.3.3 and §2.1.4.3). In this section, I discuss some of the established

methods for performing this feat. Techniques that are invoked to assess the performance of speech coding systems are described in §3.5.3. Chapter 5 presents a new technique, which draws upon some of the techniques described here. All the techniques discussed in this section operate on sampled versions of the speech signal, generating parameters that are represented as sets of discrete numbers.

There are two main approaches to encoding speech signals. One involves encoding the waveforms of signals (§3.5.1). The data rate is reduced by taking advantage of the redundancy of typical speech signals. Such schemes generally operate at medium to high data rates (16–64kbit/s) and are usually able to encode other sounds such as music or data signalling tones. In the other approach, parameters characterising models of speech signals are encoded (§3.5.2). An approximation to an original speech sound can be reconstructed from a set of parameters by means of a suitable *synthesiser*. Such schemes enable speech to be encoded at rates ranging from less than 1kbit/s to about 16kbit/s. The reconstructed speech is generally less "natural sounding" than that produced by a waveform coder. However, the reduced data rates can make such techniques attractive whenever communication or speech storage is expensive (such as in mobile telephone links or for message storage on computers). Jayant (1990) presents a recent and comprehensive review of the state of the art in high quality speech coding techniques.

3.5.1 Waveform coders

The most straightforward method of digitally encoding speech signals is to simply encode each sample as a digital word (§1.3.4). This is inefficient, however, because there is a great deal of redundancy inherent in speech waveforms. Hence, the entropy of the signal (§1.1.5) is less than the number of bits required to straightforwardly represent each sample. A multitude of techniques have been developed to take advantage of this redundancy, so as to encode the speech signals at data rates closer to their actual information content (cf. Flanagan, 1972, §1.3).

The first type of redundancy arises because of the large dynamic range of a typical speech utterance. There tends to be a wide variation in sound levels between quiet and loud segments of speech. However, the quantisation noise level is constant, depending only on the number of bits used to encode each sample. So, the signal-to-quantisation noise ratio (SNR) is greater for the high amplitude parts of speech than for the low amplitude parts. In order to encode the speech signal so that the SNR is relatively constant over a wide range of signal amplitudes, it is quantised with a *non-uniform* quantiser (Rabiner and Schafer, 1978, §5.3). This encodes the smaller amplitude samples more accurately than the larger amplitude samples. "Standard" PCM (pulse coded modulation) employs a logarithmic-based non-uniform quantiser, 8 bit words for each sample and a sampling rate of 8kHz (Jayant and Noll, 1984, Chapter 5). The data rate is therefore 64kbit/s. Because of the non-uniform quantisation, the SNR is relatively constant, at about 40dB, for a wide range of speech loudness levels (Jayant and Noll, 1984, §5.3).

In addition to the non-uniform quantisation referred to above, the quantisation process can be optimally adapted to the particular characteristics of each segment of speech. §3.5.1.1 discusses the various techniques used in these types of coders.

A third reason why redundancy exists in the sampled speech signal is that, although a sampling rate of 8kHz is required to represent all the perceptually important features of speech, many segments of speech (in particular the voiced segments) contain only insignificant spectral components above 2kHz. In addition, the higher frequency components of a speech signal can suffer higher levels of distortion before the noise becomes perceptually noticeable (due to the effects of masking, see §2.2.2.2). Hence,

a speech waveform can be effectively encoded by dividing it into frequency sub-bands and treating each one separately (§3.5.1.2).

A comprehensive treatment of all aspects of waveform coding of speech signals is provided in the recent text by Jayant and Noll (1984). Other references are cited at the relevant places throughout this section.

3.5.1.1 Adaptive coding techniques

Adaptive coding techniques take advantage of the short-term steady-state nature of speech sounds. For example, a fricative sound consists of low energy, high frequency, "noise", while a vowel sound is of much larger amplitude, is of lower frequency, and is strongly periodic. Two methods of adaptation to the short-term nature of a speech sound are employed. One technique is to adapt the quantisation step size to the expected sample amplitudes. The step size can either be calculated from the variance of the input speech, in which case it must be transmitted separately to the receiver, or it can be calculated by feedback from the quantised signal itself (Rabiner and Schafer, 1978, §5.4). The latter method is most often employed, because the quantisation information is then inherent within the encoded data stream and no "side channel" is required. However, such schemes have a greater sensitivity to transmission errors. Adaptive quantising schemes can provide an improvement of about 4–7 dB in SNR (equivalent to a one bit per sample reduction in word size for the same SNR) over standard PCM (Rabiner and Schafer, 1978, p207).

Another short-term adaptive technique involves predicting the value of the current sample from previous samples, encoding only the difference between the actual and predicted values. Because of the redundancy in speech signals, the variance of the difference is smaller than the variance of the actual speech, and fewer bits are required to represent each sample (Rabiner and Schafer, 1978, §5.5). A single predictor (equivalent to transmitting the weighted difference between successive samples) affords a SNR improvement of 4–11 dB (a reduction of about 1 bit per sample) over standard PCM (Noll, 1975).

In order to obtain further improvements in coding efficiency, the predictor coefficients can be adapted to take advantage of the short-term characteristics of the speech signal (Atal and Schroeder, 1970). The coefficients are calculated for each segment of speech by the method described in §3.2.1, with the number of coefficients normally set to about 4–10 (Rabiner and Schafer, 1978, pp229–234). Because the predictor coefficients are so effective for predicting the value of each speech sample, the *residue* can be represented with one or two bits per sample (Atal and Schroeder (1970) used one bit). Such adaptive predictive coding (APC) schemes provide SNR improvements of about 14 dB (or a reduction of 2 bits per sample) over standard PCM (Noll, 1975).

APC schemes are equivalent to the LPC vocoder schemes discussed in §3.5.2.2 with the residue (error) signal being transmitted instead of the pitch, voiced/unvoiced, and loudness parameters. After linear prediction has been performed on the speech, the residue signal contains some long-term redundancy (such as the pitch periodicity) that the short-term nature of the predictor cannot model. Hence, several schemes employ a long-term *pitch predictor*, in addition to the short-term predictor, to try and reduce the data rate required to code the residue (cf. Ramachandran and Kabal, 1989). Further refinements in predictive coding are discussed in the section on LPC vocoders (§3.5.2.2).

Predictive coders are commonly combined with adaptive quantising schemes in order to take advantage of the benefits of both. Adaptive quantisation applied to a single predictor coder (termed ADPCM, adaptive differential PCM) has become an

international standard, with a data rate of 32kbit/s providing speech quality equivalent to 64kbit/s PCM (Dècina and Modena, 1988).

3.5.1.2 Sub-band coding techniques

The redundancy in speech signals, referred to in §3.5.1.1 as a high correlation between closely spaced speech samples in the time domain, can also be viewed in a similar way in the frequency domain. Typically, a short segment of a voiced speech signal has very little energy at frequencies above 3kHz, while energy below that frequency is concentrated in the fundamental component and in the formants (cf. §2.1.4.2). By contrast, the energy in segments of unvoiced speech is typically concentrated in frequencies higher than 2kHz. In addition to these short-term concentrations of energy in relatively narrow frequency bands, the long-term spectral content of speech sounds is mainly concentrated at low frequencies. To take advantage of these characteristics of speech signals, so that they can be coded more efficiently, it is convenient to divide them into several *sub-bands*, with each being encoded separately (using any of the techniques discussed in §3.5.1.1, but often ADPCM) Typically, such a *sub-band coder* (SBC) divides the speech signal into 4–10 sub-bands. Good quality speech can be produced by encoders working at data rates down to about 24kbit/s (Daumer, 1982). Usually, the low frequency sub-bands are narrower than the higher frequency bands, and more bits are allocated to encoding the lower bands (Tribolet and Crochiere, 1979).

Another frequency-based coding technique is *adaptive transform coding* (ATC). In ATC, the speech signal is divided into segments of (typically) 10-25 ms duration, the (Fourier or other) transform of each segment is obtained, and the transformed signal coefficients are encoded for transmission (Tribolet and Crochiere, 1979). The coefficients can be encoded by techniques similar to those invoked for SBC (using fewer bits for higher frequency coefficients), vector quantisation (Chang *et al.*, 1987), or by merely encoding the spectral peaks (Almeida and Tribolet, 1984; McAulay and Quatieri, 1986). In addition, the *transform vector* for each segment can be encoded by an adaptive technique to take advantage of time variations in the spectral content of the speech signal. ATC techniques can be used to encode speech at data rates down to 16kbit/s with little distortion (Crochiere, 1978; Tribolet and Crochiere, 1979).

An important aspect of frequency-based coding schemes is that they allow straightforward implementation of techniques for *spectrally weighting* the quantisation noise. This involves forcing the spectral shape of the quantisation noise to be similar to the spectral shape of the speech signal, but offset in power level by the SNR of the system. Because of the effect of auditory masking (which renders noise inaudible if it is more than some 24dB lower in amplitude than a tone within the same critical band, as described in §2.2.2.2), spectral weighting of the noise effectively improves the perceived quality of speech produced by the coder (cf. Schroeder *et al.*, 1979). Spectral weighting can also be implemented with other coding techniques, but not in such a direct way (cf. §3.5.2.4; Atal and Schroeder, 1979).

3.5.2 Model based coders

According to the speech production model presented in §2.3.1, speech is generated by a mechanism that is controlled by (relatively) slowly varying parameters. Therefore, the speech can be encoded at low data rates if the parameters for each particular segment of speech can be easily extracted from the speech signal, encoded, and used to control a *synthesis* model which is able to regenerate a close replica of the original speech (Schroeder, 1966). A speech coder of this type is commonly called a *vocoder* (voice coder), after the introduction of the first device with this name by Dudley (1939).

The simplest speech model is the source-filter model (§2.3.1.4). Parameters for this model are the source type (voiced or unvoiced), pitch frequency (for voiced segments), source amplitude, and filter characteristics. Over the last five decades, many techniques have been developed to represent the source and filter characteristics. The aim of all these techniques is to represent the parameters in an efficient and accurate manner. §3.5.2.1 describes several of these techniques that operate with a spectral representation of the parameters, while §3.5.2.2 describes the use of LPC to represent the filter parameters. Some of the techniques that have been developed to improve the quality of LPC-encoded synthetic speech are discussed in §3.5.2.3 and §3.5.2.4.

3.5.2.1 Spectral-based vocoders

The oldest form of the vocoder is the *channel coder* (Dudley, 1939), where the spectral magnitude of the speech filter is described by the output of (typically) 10 to 20 narrow band filters. The signal from each filter is rectified and low-pass filtered to typically 20Hz for transmission. At the receiver, each one is used to control the gain of one of a similar set of band-pass filters. The excitation for the filters comes from a pulse or noise generator, controlled by pitch and VUV information from the transmitter (Schroeder, 1966). Such vocoders are able to encode speech at rates of about 2kbit/s, although the quality is generally poor at such data rates (Schroeder, 1966).

A refinement of the channel vocoder involves transmitting only the spectral peaks (McAulay and Quatieri, 1986) or formant frequencies (Linggard, 1985, §4.3). Both the formant amplitudes and frequencies must be encoded, but only the first three or four formants are required to represent the character of different speech sounds (§2.1.4). The formant frequencies can be extracted by any of the methods discussed in §3.3.3.

The extraction of only the spectral peaks (ignoring whether or not they are "formants") is a useful technique (McAulay and Quatieri, 1986) because the synthetic speech can be reconstructed from an ensemble of sinusoids at these frequencies. The phase of each sinusoid is also required for this technique, which means that higher data rates result. Such coders approach the quality and complexity of ATC techniques (§3.5.1.2).

Other frequency domain based vocoder techniques are the *phase vocoder*, where the magnitude and phase derivative of the signal in each frequency band is transmitted (Flanagan and Golden, 1966), and the *homomorphic* vocoder, where the few low-order cepstral coefficients (§3.3.2) are used to represent the spectral shape of the speech signal (Oppenheim, 1969).

3.5.2.2 LPC-based vocoders

Linear prediction (§3.2) has been employed for low data rate speech encoding since it was first applied to speech signals (Atal and Schroeder, 1968; Itakura and Saito, 1968). Because the speech filter coefficients vary only relatively slowly, they only need to be transmitted at rates of about 20–100 times per second. In addition to these filter coefficients, information on the excitation type (noise or impulse train, §3.1.2), amplitude (short-term energy, §3.1.1) and pitch period (§3.1.3) is required by the synthesiser. The data rate of the encoded speech depends on the rate at which these various parameters are updated and on the number of bits employed to encode each set.

Quantising the predictor coefficients can lead to the prediction filter becoming unstable and therefore the coefficients are usually transformed into another representation before quantising (§3.2.3). The poles of the filter are sometimes employed (Atal and Hanauer, 1971) because the filter is guaranteed to be stable if all the poles are

inside the unit circle (which can be easily ensured). However, a more usual course is to encode the reflection coefficients, since stability is ensured if these all have a magnitude less than unity (Markel and Gray, 1976, p229). In order to improve the encoding efficiency, techniques such as logarithmic compression of the coefficients (Markel and Gray, 1976, p234) or vector quantisation of each set of coefficients (Wong *et al.*, 1982) have been employed. Such approaches have led to techniques of encoding speech at data rates down to 800 bits/s (Wong *et al.*, 1982) (typically, the simple LPC techniques can encode speech at 2.4kbit/s). Readers are referred to texts such as those by Markel and Gray (1976) and Rabiner and Schafer (1978) for further details on these techniques.

3.5.2.3 Optimised excitation techniques for improved LPC-encoded speech

Speech produced by a simple source-filter synthesiser from a pitch-pulse/noise excitation signal often has an unnatural "buzzy" character. In order to improve the "quality" of the re-synthesised speech, various techniques are employed to obtain more "optimal" excitation signals.

Improvements are obtained most readily by employing "shaped" pulses rather than simple impulses for the voiced excitation (Holmes, 1973). Often a triangular shaped pulse is employed, since this has a spectrum that is similar to that of the glottal waveform (Flanagan, 1972, §6.241). Another technique employs an excitation that differs in each of several frequency sub-bands (Fujimura, 1968; Griffin and Lim, 1988). Such an excitation takes into account the occurrence of "mixed" excitation (see §2.2.1). Also, because voiced speech is not perfectly periodic, its spectrum, especially at higher frequencies, is not composed entirely of harmonics of the "fundamental". Employing an excitation signal that differs in several frequency sub-bands can therefore help to produce less "buzzy" sounding speech. The reconstructed speech quality can also be improved by adjusting the phase of the excitation spectrum so that the excitation is not so "peaky" (cf. Schroeder, 1975). In addition, a small random "jitter" of the phase can help to introduce (more natural sounding) pitch perturbations into the reconstructed speech signal (Fujimura, 1968; Kang and Everett, 1985).

Another method of improving the excitation source is to encode the LPC residue. This is called *residue excited LPC* (RELP) coding and is very similar to the APC techniques discussed in §3.5.1.1. In traditional RELP coding, the excitation is derived from a low-pass filtered version of the residue signal (Markel and Gray, 1976, §10.4). In order to reduce the data rate required, techniques such as sub-band coding of the residue (Un and Lee, 1984) have been employed. For voiced speech, the residue exhibits a pronounced periodicity at the pitch frequency (§3.1.3). This can be removed with a *pitch prediction filter* (Ramachandran and Kabal, 1989). Apart from this, RELP techniques, and the methods discussed in subsequent paragraphs, typically do not require the use of VUV or pitch analysis, because this is inherent in the excitation coding.

Techniques for modelling the LPC residue can be invoked to parametrically represent the "optimum" excitation signal at low data rates. For example, Sreenivas (1988) computes "spike" and "noise" components that optimally model the residue signal. In another approach, Atal and Remde (1982) model the residue by a few (typically 4–10) discrete pulses in each segment. The amplitudes and positions of each pulse are calculated in order to minimise the weighted error between the original and the reconstructed speech (Atal, 1985). Methods for constructing these pulses are described in more detail in §3.5.2.4.

Yet another approach to modelling the excitation is to produce a code book of possible excitation sequences, choosing that sequence which minimises the resultant reconstruction error (Atal and Schroeder, 1984). A code book of 1024 different ex-

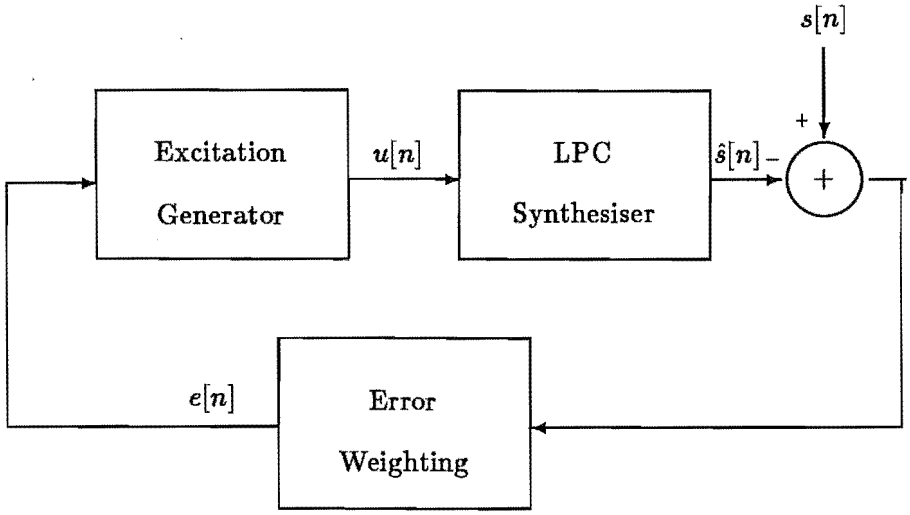


Figure 3.9. Analysis-by-synthesis technique for obtaining an optimal excitation sequence. The perceptual weighting increases the amount of allowable error in spectral regions where the speech energy is high.

citation sequences (inferring a code of 10 bits per frame) is typically employed, with each sequence being about 5ms in duration (Atal and Schroeder, 1984; Davidson and Gersho, 1986).

3.5.2.4 Multi-pulse excitation technique

The multi-pulse method of obtaining an "optimal" excitation sequence for LPC-encoded speech is described in more detail than the other methods mentioned in §3.5.2.3 because of its similarity to the new technique that I present in Chapter 5.

Multi-pulse (MP) excitation models the excitation signal as a few discrete pulses. Typically, 4–10 pulses are employed per 10ms segment. MP-LPC is often termed an *analysis-by-synthesis* technique. In contrast to the RELP techniques described in §3.5.2.3, for which the excitation is (usually) obtained from the residue after performing LPC analysis (note, however, that the CELP technique described in §3.5.2.3 is also an analysis-by-synthesis scheme), the excitation signal for analysis-by-synthesis schemes is obtained in an iterative fashion by minimising the error between the original speech and the speech that would be reconstructed by the trial excitation (cf. Kroon and Deprettere, 1988). Fig.3.9 shows a block diagram of an analysis-by-synthesis scheme. The error is weighted so as to reduce its value in regions of the spectrum where it is perceptually less noticeable (Atal and Schroeder, 1979). A simple weighting function $W(z)$, described by Atal and Remde (1982), is formed from the LPC filter $A(z)$ by

$$W(z) = \frac{A(z)}{A(z/\gamma)}, \tag{3.32}$$

where γ is a constant between 0 and 1 that determines the amount of de-emphasis given to the formant peaks by $W(z)$.

For the case of the multi-pulse LPC (MP-LPC) technique, the excitation generator identified in Fig.3.9 determines the positions and amplitudes of the pulses that comprise the excitation signal (Atal and Remde, 1982). In this section I briefly describe the method by which these pulses are located. Further details are presented by Atal (1985) and Singhal and Atal (1989).

The perceptual weighting of the error signal that is indicated in Fig.3.9 can be applied equivalently to the actual and trial speech signals. The weighted mean square error over the interval $0 \leq n < N$ then becomes

$$E = \sum_{n=0}^{N-1} [y[n] - \hat{y}[n]]^2, \tag{3.33}$$

where $y[n]$ is the weighted speech signal and $\hat{y}[n]$ is the weighted synthetic speech. The synthetic speech is produced by a convolution between the weighted impulse response $h[n]$ of the LPC synthesis filter $A(z)$ and the trial excitation

$$u[n] = \sum_{k=0}^{m-1} \beta_k \delta[n - n_k], \tag{3.34}$$

where β_k and n_k are the amplitude and position respectively of the k^{th} excitation pulse respectively. The synthetic speech can therefore be expressed as

$$\hat{y}[n] = \sum_{k=0}^{m-1} \beta_k h[n - n_k] + \hat{y}_0[n], \quad 0 \leq n < N, \tag{3.35}$$

where $\hat{y}_0[n]$ is the memory of the synthesis filter from the previous frame. $\hat{y}_0[n]$ can be computed (assuming that the filter $h[n]$ does not change too much between each frame) by means of

$$\hat{y}_0[n] = \sum_{k=0}^{m-1} \hat{\beta}_k h[n - \hat{n}_k], \quad 0 \leq n < N, \tag{3.36}$$

where $\hat{\beta}_k$ and $\hat{n}_k$ are the amplitudes and positions of the m excitation pulses in the previous frame.

Substituting (3.35) into (3.33), setting the derivative of E with respect to each of the pulse amplitudes β_k to zero, and rearranging, leads to the set of equations

$$\sum_{k=0}^{m-1} \beta_k \alpha_{n_k n_j} = c_{n_j}, \quad 0 \leq j < m \tag{3.37}$$

where α_{ij}, which is the autocorrelation of $h[n]$, is defined to be

$$\alpha_{ij} = \sum_{n=0}^{N-1} h[n - i]h[n - j], \quad i, j = 0 \ldots N - 1, \tag{3.38}$$

and the vector c_j is the cross-correlation of $h[n]$ and the signal $\bar{y}[n] = y[n] - \hat{y}_0[n]$:

$$c_i = \sum_{n=0}^{N-1} \bar{y}[n]h[n - i], \quad i = 0 \ldots N - 1. \tag{3.39}$$

Substituting (3.35) and (3.37) into (3.33) leads, after some manipulation, to the expression

$$E_{\min} = \sum_{n=0}^{N-1} \bar{y}^2[n] - \sum_{k=0}^{m-1} \beta_k c_{n_k} \tag{3.40}$$

for the minimum value of E. Since only the second term in RHS (3.40) depends on the excitation, the optimum value for $E_{\min}$ can be found by finding a solution (3.37) which maximises this term. This is computationally impractical in general, and so sub-optimal schemes for locating the pulses are invoked (Singhal and Atal, 1989). One

such scheme involves locating the pulses sequentially in a recursive procedure. If $u[n]$ consists of only one pulse, (3.37) reduces to

$$\beta_0[i] = c_i/\alpha_{ii} \tag{3.41}$$

and (3.40) to

$$E_{\min}[i] = \sum_{n=0}^{N-1} \bar{y}^2[n] - c_i^2/\alpha_{ii}, \tag{3.42}$$

where $\beta_0[i]$ and $E_{\min}[i]$ are the values of β_0 and $E_{\min}$ respectively when the pulse is located at $n = i$. The error is minimised when the second term in RHS (3.42) is at a maximum. Hence, the pulse is located at the position $n_0 = i$ for which c_i^2/α_{ii} is maximised.

The second pulse of $u[n]$ can be found by setting $m = 2$ and incorporating β_0 and n_0, found as described in the previous paragraph, into (3.37) and (3.40). The recursive procedure for finding the j^{th} pulse, for $j = 1 \ldots m-1$, is:

$$c_i^{(j)} = \left[c_i^{(j-1)} - \beta_{j-1}\alpha_{n_{j-1}i}\right] \quad i = 0 \ldots N-1., \tag{3.43}$$

$$n_j = \arg\max_i \left[c_i^{(j)}\right]^2 /\alpha_{ii}, \tag{3.44}$$

and

$$\beta_j = c_{n_j}^{(j)}/\alpha_{n_j n_j}. \tag{3.45}$$

After all m pulses have been found by this procedure, their amplitudes are re-optimised by solving (3.37). Fig.3.10 shows a segment of speech, the MP excitation sequence (found as described above) and the resultant synthetic speech waveform. Eight pulses were obtained in each segment (each of which was 7.5ms in duration), with the spacing between adjacent segments being 5ms. The pulses in the overlap regions between segments were discarded as suggested by Singhal and Atal (1989). The LPC filter was composed of 10 coefficients obtained at 7.5ms intervals. Each LPC filter was computed from a segment of 10ms in duration by the autocorrelation method (§3.2).

The method of locating the pulses described in the above two paragraphs does not work well when m is large. This is because the pulse amplitudes remain fixed after they have been initially located (Lefevre and Passien, 1985). The first few pulses may therefore be inaccurate because the effect of later pulses has not been accounted for. The pulse amplitudes can be re-optimised by application of (3.37) after each pulse has been found, but this increases the amount of computation required. Singhal and Atal (1989) present an "optimised" procedure, in which the pulse amplitudes are implicitly allowed to vary until all the pulses have been located.

Further improvements to MP-LPC include the use of pitch prediction to capitalise on the "periodicity" of the pulse signal (cf. Araseki *et al.*, 1986; Singhal and Atal, 1989), re-computation of the LPC filter parameters by taking into account the newly-found MP excitation signal (Singhal and Atal, 1983), and various "optimal" methods of locating the pulses (cf. Lefevre and Passien, 1985; Boyd, 1987).

Another variant of the multi-pulse approach, which has recently gained prominence as the coding technique adopted for the European digital mobile radio scheme (Natvig, 1988), is called *regular pulse excitation* (RPE). In this technique, the pulses are positioned at regular spacings (e.g. every five samples) throughout each segment. Hence only the position of the first pulse needs to be specified, together with the amplitudes of each of the pulses (Kroon *et al.*, 1986). The RPE scheme typically requires more pulses than MP-LPC in order to attain the same performance, but this is more than compensated for by only having to specify one position (Kroon and Deprettere, 1988).

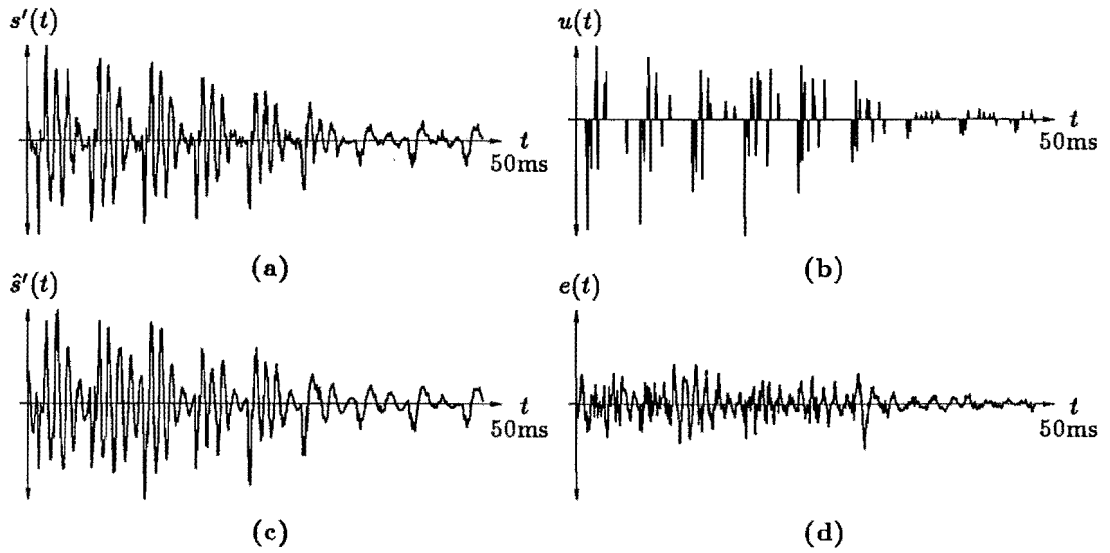


Figure 3.10. Example of multi-pulse coding of speech. **a:** (Differentiated) speech waveform (taken from utterance AM-RAIN1). **b:** Multi-pulse excitation sequence. **c:** Synthetic speech waveform. **d:** Error signal.

3.5.3 Performance evaluation of speech coding techniques

The evaluation of the performance of a particular speech encoding scheme requires that the "distortion" of the speech signal by the system be quantified in some meaningful manner. In this section I describe some of the techniques that have been developed to assess the *quality* of speech produced by low data rate speech coding schemes. §3.5.3.1 is concerned with subjective measures of speech quality, as judged by human listeners, while §3.5.3.2 describes techniques for obtaining "objective" measures of the amount of distortion suffered by a processed speech signal.

The quality of speech produced by a particular encoding system is enumerated by the quality "score" or "rating" of utterances processed by the system. This score may be obtained by the subjective assessment methods described in §3.5.3.1, in which case it is a comparative rating on a perceptual scale of "speech quality", or by the objective assessment methods described in §3.5.3.2, in which case it is a rating on a specified scale of "speech fidelity". The different quality scales can be related to each other by comparing the scores for particular utterances when assessed by each of the different methods (see §3.5.3.1).

Any measure of the "performance" of a speech encoding scheme must of course include reference to the data rate of the encoded speech, the computational complexity of the requisite processing, and the sensitivity of the scheme to errors in the encoded data (cf. Natvig, 1988). The cost of data storage or transmission is the main reason for utilising low data rate speech encoding schemes. The trade-off between data rate and speech quality can be determined by evaluating the quality scores of utterances encoded at a range of data rates. The relative performances of different schemes at any particular data rate can then be compared by examining their respective quality scores. The scores should also be compared at a range of transmission error rates, because of the trade-off between error rate and data rate on any data transmission channel (cf. Hamming, 1980). The computational complexity of an encoding scheme is an important consideration because it directly influences the cost of the processing hardware required to implement it.

3.5.3.1 Subjective measures of speech quality

There are several relevant aspects of the "goodness" of speech produced by a speech encoding or synthesis scheme. *Intelligibility* is the accuracy with which words and syllables can be correctly identified after they have been processed by the system. Another aspect is the *naturalness* of the speech, which for speech encoding schemes is a measure of how well the speech mimics the voice characteristics of the original speaker. Naturalness is of particular concern for higher data rate schemes, whereas intelligibility ratings are of more importance for low data rate schemes (Kitawaki and Nagabuchi, 1988). In this section I describe methods of evaluating the perceived *quality* of processed speech, where quality is a term that encompasses all aspects of "how good" a certain speech signal sounds to human listeners. It therefore includes both the intelligibility and naturalness of the speech. Techniques for assessing the intelligibility of speech on its own (cf. Fairbanks, 1958; Kitawaki and Nagabuchi, 1988) are neglected here.

There are two main methods of evaluating the perceptual quality of speech. Both of them involve playing speech utterances to groups of listeners and obtaining their opinions about their perception of the quality of each utterance. Before reviewing these assessment methods, I discuss, in the next two paragraphs, criteria for choosing utterances for the evaluation.

The utterances employed in the evaluation should be of phrases that are phonetically balanced, so that they cover a range of sounds. A list of appropriate phrases has been published (cf. IEEE, 1969)). Alternatively, a variety of "phoneme-specific" sentences can be employed to evaluate the ability of the processing system to reproduce different "types" of phonemes, such as vowels, fricatives, or nasals. Huggins and Nickerson (1985) list suitable phrases and describe procedures for evaluating the results of such an approach. The content of each phrase should be neutral so that the listeners can focus on the quality of the utterances more easily. It is usual to have several different phrases of similar style and length (IEEE, 1969). Utterances spoken by a range of speakers may be employed in the assessment so as to avoid biasing the results by the peculiarities of any particular speaker's voice quality (IEEE, 1969).

In each of the assessment methods, the *test* utterances, which are utterances that have been processed by the systems under investigation, are augmented by *reference* utterances, which are utterances that have been distorted by a quantified level of noise. Hence the subjective quality ratings can be associated with equivalent signal-to-noise ratios. Reference utterances are obtained by corrupting utterances, that in other respects are the same as the ones used for the test utterances, with various levels of noise according to the formula

$$r(t) = s(t) + kn(t) \tag{3.46}$$

where $r(t)$, $s(t)$, and $n(t)$ represent the reference, "perfect" speech, and noise signals respectively, and k represents the level of distortion. (IEEE, 1969). A convenient way to obtain a noise signal that is uncorrelated with $s(t)$, but at the same time is related to the short-term amplitude of $s(t)$ (so that the SNR is relatively constant for different levels of speech amplitude) is to multiply $s(t)$ by a spectrally weighted random noise signal $n_o(t)$:

$$n(t) = s(t)n_o(t) \tag{3.47}$$

where $n_o(t)$ is obtained by filtering white noise through a filter with a slope of -10dB/decade (IEEE, 1969; Rothauser *et al.*, 1968), so as to match the average spectral content of typical speech signals. The SNR of the reference signal defined by (3.46) and (3.47) is given by

$$\mathrm{SNR}_{\mathrm{ref}} = 20 \log \left\{ \frac{1}{kR_{n_o}} \right\} \tag{3.48}$$

where R_{n_o} is the RMS level $n_o(t)$. Further aspects of generating reference utterances are discussed by Rothauser *et al.* (1968) and the authors of (IEEE, 1969). An alternative type of reference signal, in which $n_o(t)$ is replaced by a signal that has unit amplitude but whose phase changes randomly by π, is described by Schroeder (1968).

The *category judgement* method of quality assessment involves asking each listener to rate each utterance separately on a scale of "quality", usually from one to five (cf. IEEE, 1969; Rothauser *et al.*, 1971; Huggins and Nickerson, 1985; Daumer, 1982; Kitawaki and Nagabuchi, 1988). The scale is usually labelled: 1 — unacceptable; 2—poor; 3—fair; 4—good; and 5—excellent (IEEE, 1969). The scores for each utterance are averaged over the listeners to produce a *mean opinion score* (MOS) for that utterance.

The category judgement test requires that the listeners have an idea of what the different categories mean in terms of speech quality. This can be effected by presenting them with utterances of a specified quality, such as those at the extremes, before they begin the test (IEEE, 1969). This "anchoring" can also be augmented by presenting examples of the utterances from the whole range of qualities that are to be tested. This gives the listeners an opportunity to become familiar with the material, and enables them to form opinions about the different categories. A subset of the utterances is often repeated, say at the beginning and the end of the test, to examine the consistency of each person's scoring (Huggins and Nickerson, 1985)

In *comparison* or *ranking* methods of quality assessment, each listener is asked to compare pairs or groups of utterances, and rank them accordingly. There are several approaches to accomplishing this. The *paired comparison*, or *iso-preference*, test requires that the listener judge which is the better of every pair of utterances (cf. Rothauser *et al.*, 1968; IEEE, 1969; Huggins and Nickerson, 1985). The probability matrix of preference between each pair of utterances is then obtained by averaging the responses from all listeners. Pairs of speech utterances are deemed to be equal in quality when they are preferred with equal probability (hence the term iso-preference). The utterances can therefore be ranked in order of quality.

The paired comparison test is very time consuming, since about $O(N^2)$ comparisons must be made (where N is the number of utterances). One way of reducing the number of comparisons is to implement an *elimination tournament* strategy, where utterances are "eliminated" from the "tournament" when they "lose" a certain number of comparison tests. The ranking of a particular utterance is the number of comparisons that it "wins". Rosenberg (1971) describes the details of this type of test. Another technique for comparing many utterances is to present them in small groups, asking each listener to rank the utterances within the group (cf. IEEE, 1969; Rothauser *et al.*, 1971). Such a test requires fewer comparisons than the iso-preference method when the total number of utterances is large.

Each of the quality assessment methods discussed above produces a rating for a particular utterance (and hence for the coding scheme that produced that utterance) on an arbitrary scale. The scales for the different methods can be related by performing different tests on the same set of utterances. Huggins and Nickerson (1985) assert that the category judgement and iso-preference methods produce virtually identical results. However, Rothauser *et al.* (1971) find that the scales produced by the category judgement and iso-preference methods cannot be directly related. They postulate that this is because the five quality steps in the category judgement method are too coarse So, they propose a modification of this method that allows listeners to rate each utterance on a decimal (non-quantised) scale from 0 to 10. Their modified method produces results that are more in accord with those from the iso-preference method.

The scale for any method can be related to an SNR scale by incorporating

reference signals into the test (cf. Rothauser *et al.*, 1971; Kitawaki *et al.*, 1988). However, "quality" is in general not a uni-dimensional quantity, since it is related to several aspects of a speech signal, such as its naturalness, the amount of additive noise, or the accuracy with which phonetic features are represented. Statistical techniques such as multidimensional scaling can be invoked to examine the relationship between the quality score and its various physical or perceptual correlates (cf. Hecker and Guttman, 1967; Rosenberg, 1971; Cummiskey *et al.*, 1973; Huggins and Nickerson, 1985).

Goodman and Nash (1982) compare the results of subjective measures (via category judgement tests) of the quality of each of several speech transmission systems obtained from listeners in seven different countries. They find that the only significant differences in scores between countries occurs for listeners in Japan (lower MOS ratings than the all-country average) and the USA (higher MOS ratings).

Further details of the methodology of these tests, and a comprehensive review of previous work on subjective speech quality evaluation can be found in *IEEE Recommended practice for speech quality measurement* (IEEE, 1969).

3.5.3.2 Objective measures of speech quality

By an "objective" measure of speech quality I mean a rating that can be numerically computed from the speech signal. Such measures are necessary adjuncts to subjective measures (§3.5.3.1), both because of the time consuming nature of performing subjective tests, and because of the difficulty of ensuring conformity between subjective judgements made under different conditions (especially those made in different laboratories). Of course, as pointed out by Takahashi (1988), a subjective evaluation of quality measures the opinions of the users directly, which is of prime importance in any commercial environment. In this section I describe some of the objective measures that have been proposed to measure the "quality" of processed speech signals. More properly, they measure the *distortion* between the processed and unprocessed signals. The quality can be expressed as the proportional amount of distortion in the processed signal.

Perhaps the simplest measure of quality is the SNR, which is defined by

$$\mathrm{SNR} = 10 \log \frac{\langle y[n]^2 \rangle_n}{\langle \{y[n] - x[n]\}^2 \rangle_n} \tag{3.49}$$

where $x[n]$ and $y[n]$ are the input and output signals respectively of the speech processing system under test (cf. Dimolitsas, 1989). The SNR of an utterance is usually defined as the average of the ensemble of SNR values computed, by means of (3.49), from short sequential segments of the utterance. The SNR defined in this way has two deficiencies. The first is that the *relative* noise level during the "silent" segments of speech can be high, and hence the SNR can be large and negative, without the noise being perceptually noticeable (because the signal energy is so low anyway). The *segmental SNR* is defined so that each segment has a minimum SNR of 0dB. For the m^{th} segment (of duration M samples), the segmental SNR $L_s[m]$ is defined by (Mermelstein, 1979)

$$L_s[m] = \log \left\{ 1 + \frac{\sum_{n=mM}^{(m+1)M} x[n]^2}{\sum_{n=mM}^{(m+1)M} \{y[n] - x[n]\}^2} \right\}. \tag{3.50}$$

The average segmental SNR for the utterance $\mathrm{SNR}_{\mathrm{seg}}$ is then defined as

$$\mathrm{SNR}_{\mathrm{seg}} = 10 \log \{ \exp(\langle L_s[m] \rangle_m - 1) \}. \tag{3.51}$$

Crochiere (1978) defines an alternative form of segmental SNR in which the SNR of each segment is computed by (3.49), but only the values for segments whose energy is greater than some threshold are averaged to give the actual SNR value.

The other deficiency of SNR ratings, which applies to both the simple and the segmental SNR measures defined in the previous paragraph, is that many waveform distortions are not perceptually relevant. For instance, sign inversion of a speech signal results in a SNR of -6dB, yet the speech sounds the same as non-inverted speech. In addition, noise that is correlated with the signal generally has a lower perceptual effect than uncorrelated noise (Jayant, 1974). Most vocoder-type speech coding techniques do not preserve the waveform, yet they produce perceptually adequate speech. Several *distance measures* or *distortion measures* have been developed to characterise the perceptually relevant distortions of a speech signal (cf. Atal and Schroeder, 1979; Schroeder *et al.*, 1979; Kitawaki *et al.*, 1988; Dimolitsas, 1989). These are all based on short-term spectral representations of the speech sounds. They are closely related to the distance measures employed in speech recognition schemes (§3.6.1).

The most straightforward spectral distance measure is the difference between the short-term spectral, cepstral, or LPC coefficients of the processed and unprocessed speech signals (cf. Kitawaki *et al.*, 1988; Dimolitsas, 1989). Kitawaki *et al.* (1988) define a cepstral distance d_{CEP} measure as

$$d_{CEP} = 10 \log \sqrt{2 \sum_{i=1}^{P} \{c_i^x - c_i^y\}^2} \tag{3.52}$$

where c_i^x and c_i^y are the (P^{th} order) cepstral coefficients (§3.3.2) of the unprocessed and processed speech signals respectively. The average distance measure for an utterance is obtained by averaging the value of d_{CEP} computed from each short segment of speech. Kitawaki *et al.* (1988) relate this measure to the perceptual MOS scale of speech quality by comparing the results obtained in a subjective and objective evaluation of a set of speech utterances. They find that the cepstral distance gives results with a good correspondence to those obtained by the category judgement method (§3.5.3.1). The short term log spectral distance d_{sp} is defined by

$$d_{sp_m} = \langle \log |X_m(f)|^2 - \log |Y_m(f)|^2 \rangle_f \tag{3.53}$$

where $X_m(f)$ and $Y_m(f)$ are the spectra of the m^{th} segment of the original and processed speech signals respectively.

Other distance measures have included several commonly employed in speech recognition. Reviews of these are given by Gray and Markel (1976) and Nocerino *et al.* (1985), from the point of view of speech recognition, and Dimolitsas (1989) from a speech quality assessment point of view. Schroeder *et al.* (1979) present a comprehensive scheme for processing the short-term spectra of speech sounds according to the characteristics of human sound perception. Their scheme takes into account both the characteristics of sound-to-neural transduction and the effects of noise masking. The speech degradation is expressed as a number between zero and unity, estimating the ratio of perceived noise to perceived speech loudness.

3.6 Other applications of speech processing techniques

In this section, I discuss some of the applications, other than low data rate speech encoding, of the speech analysis techniques introduced in §3.1 through §3.4. Techniques that are commonly invoked in speech recognition schemes are described in §3.6.1, while

§3.6.2 provides an introduction to the field of speaker recognition. §3.6.3 gives a brief overview of text-to-speech conversion. Finally, §3.6.4 describes some of the analysis techniques that have been applied to speech therapy and the diagnosis of vocal disorders.

Since my research was not directly concerned with any of the applications described in this section, they are treated somewhat cursorily. However, this section is included in order to complete the coverage of speech processing techniques which is the subject of this chapter, and because of the similarities of some of the techniques to those discussed in Chapters 6 and 7. In addition, the new speech analysis techniques that I present in Chapters 4 and 5 could feasibly be applied to any of the problems described here (see the discussion in §8.2).

3.6.1 Automatic speech recognition

The area of speech recognition is one of the most challenging currently facing speech researchers. Although my research has not been concerned with speech recognition, the difficulties that are encountered, and the techniques for overcoming them, are relevant to many other problems, such as the classification of cough sounds and animal vocalisations discussed in Chapters 6 and 7 respectively. In this section I briefly review the techniques that have been developed for recognising speech. The related field of speaker recognition is similarly treated in §3.6.2.

Note that I am not concerned here with techniques of "machine understanding", but only of converting speech sounds into their text equivalent. Techniques of language understanding and "artificial intelligence" are treated in depth by many authors (cf. Woods, 1985). It should be kept in mind, however, that it will probably be necessary to incorporate "higher level" linguistic processing into any computer algorithm that will (if, in fact, this ever eventuates) be capable of recognising speech as well as humans do (see the discussion in §2.1.3.3). The types of applications for which speech recognition schemes are useful include "hands free" controllers for machines, automatic dictation machines, and telephone services such as airline reservation systems (cf. Rabiner and Levinson, 1981).

The basis of any recognition technique is to compare the *test pattern* of an (unknown) input signal with a number of standard *template patterns*. The template that best matches the input pattern is judged to represent the unknown signal. In the context of speech recognition, the signals consist of individual words or word units (such as phonemes or syllables). Speech recognition schemes can be divided into those that recognise individual words, termed *isolated word* recognition schemes, and those that connected words or sentences, which are termed *continuous*, or *connected, speech* recognition schemes (Rabiner and Levinson, 1981; Vaissière, 1985). The discussion in this section is limited to isolated word recognition, although many of the concepts are also applicable to recognition of continuous speech. Recognition schemes are termed *speaker independent* if they are designed to recognise words from any speaker and *speaker dependent* if they are limited to a single speaker (Vaissière, 1985).

Successful speech (or speaker) recognition relies on the estimation of features which adequately describe the phonetic content of speech sounds (or characteristics of an individual's voice). All of the analysis techniques described in §3.1 through §3.3 can be employed to extract features for the recognition of words and voices. In §3.6.1.1 I describe the types of features that are commonly invoked for speech recognition purposes. §3.6.1.2 introduces the techniques that are employed to match the test and reference patterns, while §3.6.1.3 briefly discusses the difficulties involved in training speech recognisers.

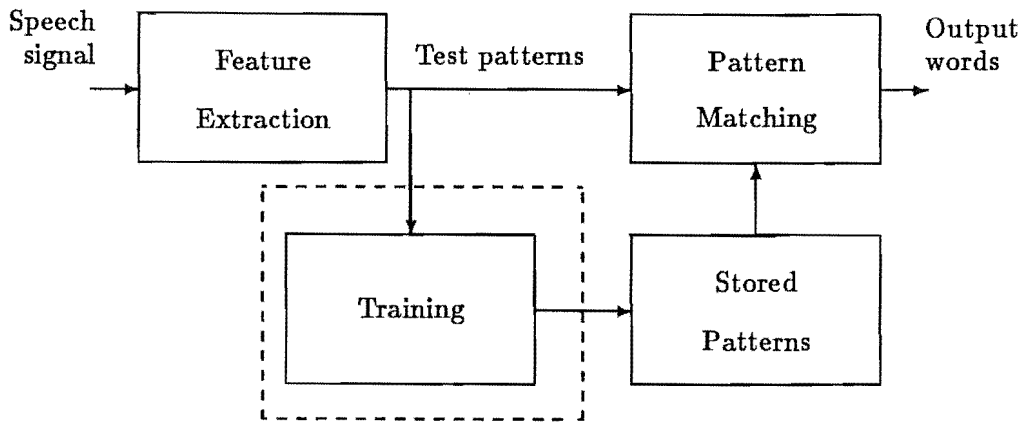


Figure 3.11. Block diagram of a speech recognition scheme

3.6.1.1 Features and distance measures

Fig.3.11 shows a block diagram of a word recognition scheme. The feature measurement stage consists of one (or more) of the techniques described in §3.1 through §3.2. Features that are commonly employed for word recognition include the short time spectral content or formants, LPC coefficients, and (low order) cepstral coefficients (Davis and Mermelstein, 1980; Rabiner and Levinson, 1981; Juang *et al.*, 1987). Cepstral coefficients currently appear to provide the best recognition performance (Rabiner *et al.*, 1989). The features are often weighted so as to account for the perceptual characteristics of human hearing. For example, Davis and Mermelstein (1980) weight the formant frequencies according to the mel-scale (§2.2.2.2) description of human speech perception. However, in their comparison of various distance measures, Nocerino *et al.* (1985) find that such weighting does not improve recognition performance. Cohen (1989) and Ghitza (1987) both process the short-term spectral content of an input sound according to the characteristics of human sound-to-neural transduction (see §3.3.3). In addition to these perceptual weightings, the different components of the feature vector are often weighted according to an experimentally obtained estimate of their individual importance in characterising the differences between sounds (Juang *et al.*, 1987; Tohkura, 1987).

Perceptual weighting of features is often implemented by weighting the *distance measure* used to characterise the similarity between pairs of features (Gray and Markel, 1976). The particular distance measure employed is important because it determines how the similarity between patterns is evaluated, and hence affects the performance of the recognition (Gray and Markel, 1976; Juang *et al.*, 1987). Most successful distance measures are based on the log likelihood ratio (Itakura and Saito, 1970; Nocerino *et al.*, 1985). This matches the spectral peaks more closely than the intervening dips (§3.2.2), which conforms to the perceptual characteristics of masking, whereby spectral prominences mask regions of lower energy (§2.2.2.2).

3.6.1.2 Pattern matching techniques

Because of the great variation in the manner in which words are uttered, sophisticated schemes for matching the input features to the templates must be employed.

Words are composed of temporal sequences of phonetic features (see §2.1.3.1). Thus, the patterns that must be compared in the second stage of the recogniser (Fig.3.11) consist of variable length lists of feature vectors. The comparison stage involves characterising the similarity between the test pattern and each of the reference

patterns. Because speaking rates vary widely, techniques for aligning the patterns must be employed. The end-points of isolated words can be detected by silence detection techniques (§3.1.2). In order to match two patterns of unequal length, techniques of non-linearly "warping" (often termed *Dynamic time-warping*, or DTW) the lengths of each pattern to obtain the optimum match are invoked (cf. Rabiner and Levinson, 1981).

The final stage of the recogniser (Fig.3.11) consists of deciding on the correct word or list of words. In some systems, the several "best" matches are chosen (Rabiner and Levinson, 1981), and further decisions are made at a higher level (perhaps by applying syntactical rules). In addition, further (more computationally expensive) techniques may be applied to the patterns if they are found to belong to certain sub-classes of similar words (Fissore *et al.*, 1989). In this way, large vocabularies can be searched relatively quickly. A preselection stage is performed, separating the words into broad phonetic groups. The word is then matched to one of the templates in the (much smaller) sub-vocabulary (Fissore *et al.*, 1989).

Other techniques for comparing patterns employ statistically-based models of the templates, with the time-variation of the features implicit in each model. One such method is *hidden Markov modelling* (HMM), in which each reference pattern is represented as a HMM (Rabiner and Juang, 1986). A HMM consists of several *hidden states*, each of which possesses an associated *observation vector* that describes the probabilities of each possible feature arising from that state. The complete HMM is characterised by this observation matrix and the matrix of *transition probabilities* between each state. Such models appear to be successful in speaker-independent and connected-speech recognition schemes (Rabiner *et al.*, 1989).

Another matching scheme makes use of *neural networks* (Lippmann, 1987). These are still at a largely experimental stage, but because they appear to model the human brain, it is hoped that they will solve many of the problems that beset traditional matching schemes (cf. Burr, 1988). A neural network consists of a large number of interconnected *nodes*. Each node has a large number of weighted inputs, and its output is similarly connected to the inputs of many other nodes. The value of the output is a non-linear function of the sum of the weighted inputs to that node. Neural networks naturally conform to a parallel architecture, with the operation of each node being particularly simple. Hence, real-time operation, even with very intricate systems, should be relatively straightforward to achieve. Several *topologies* have been investigated for organising the layout of neural networks. One common topology is the *multi-layered* approach, where interconnections are only made between nodes on different layers (Lippmann, 1987).

In a neural network, the reference patterns are implicit in the weightings applied to the inputs of each node (Waibel *et al.*, 1989). An advantage of neural networks for speech recognition is that the "decision rules" (or word models for HMM techniques) do not need to be explicitly chosen *a priori*. So, neural networks are able to optimise themselves to form arbitrarily complicated "decision boundaries" between each class of word or phoneme (Burr, 1988). Also, instead of a separate model for each word, as for the HMM technique, all words are catered for by the one network. Hence the training process can minimise the response of competing words at the same time as maximising the response of the correct word (cf. Kohonen, 1988). This can increase the level of discrimination between words (Bourlard and Wellekens, 1989).

A difficulty with applying neural networks to speech recognition is that there is no straightforward way to incorporate information about the temporal variation of the features, which is obviously of great importance for reliable speech recognition. Waibel *et al.* (1989) incorporate temporal information about the speech signal by storing several

time-delayed feature-sets and using them as additional inputs to the network.

3.6.1.3 Training methods

For any of the techniques mentioned in §3.6.1.1, the reference patterns must be obtained by a *training* process. Training consists of analysing many examples of a particular word and determining one or more reference patterns from the ensemble of examples. Techniques such as clustering are employed to find the pattern vectors that optimally represent the input words. The centroid of a cluster is therefore all that is necessary to represent all of the points in that cluster (Gray, 1984). In order to account for the variations resulting when several speakers use the system, more than one template or centroid may be stored for each word (Rabiner and Levinson, 1981).

In the HMM and neural network approaches, training involves an iterative process of (re)calculating the probability matrices and adjusting the node input weights, respectively, for each test word, in order to find an optimal solution (cf. Lippmann, 1987; Rabiner and Juang, 1986).

3.6.2 Speaker recognition techniques

Speaker recognition is a problem that is akin to speech recognition (§3.6.1) but with the patterns of words replaced by patterns representing the characteristics of individuals' voices. In §3.6.2.1 I introduce the requirements of speaker recognition schemes, while in §3.6.2.2 I briefly describe the features that are employed in such schemes.

3.6.2.1 Speaker recognition requirements

Speaker recognition schemes can be divided into two types — those concerned with *identifying* a given individual from a (known) population, and those that attempt to *verify* the claimed identity of a speaker (O'Shaughnessy, 1986). The recognition strategies differ somewhat between these two cases, although the same features (§3.6.2.2) can be employed.

In a speaker identification scheme, the system must compare the test pattern with each of the reference templates for the speakers that it knows about. A decision is then made as to the identity of the speaker. If none of the templates are "close enough", the speaker is classed as "not known".

By contrast, a speaker verification system must decide whether the voice characteristics of the speaker are "close enough" to the reference template of the speaker's claimed identity. The system must make a decision about the "closeness' of the test and reference patterns by comparing the distance between them with the expected inter-speaker and intra-speaker variation of distances (Doddington, 1985). A verification system can make two types of errors — either rejecting a "true" person, or accepting an "imposter". The two types of error are in opposition, since raising the "closeness" threshold to allow for more intra-speaker variation also increases the chances of an imposter being accepted (cf. Birnbaum *et al.*, 1986).

In the speaker verification scenario, the population size is potentially unlimited, since the system presumably does not have templates for all the *imposters* who might make a false claim. However, such a system only needs to make a "Yes"/"No" decision, while a speaker identification system must make a choice of one out of many speakers. Hence, the error rates for verification systems are usually lower than those for identification schemes (Doddington, 1985).

Speaker recognition schemes can additionally be divided into those that are *text-dependent*, and those that are *text-independent*. In the text dependent approach,

which is often employed in speaker verification schemes (O'Shaughnessy, 1986), a standard phrase is used for all training and verification. Because of this, features which relate to the actual text, such as the way in which particular sounds are pronounced, can be invoked for recognition. By contrast, text-independent approaches (which are appropriate when the co-operation of the speaker is not assured) are restricted to features which describe the "average" characteristics of a person's voice, irrespective of the actual text (Furui, 1981). Generally, recognition performance is better in the text-dependent approach, because greater control is possible over the test conditions (Doddington, 1985).

3.6.2.2 Features for speaker recognition

There are two classes of features that are employed for speaker recognition. The first relates to the "average" characteristics of a person's voice, while the second contains information about the *dynamic* or temporal structure of how they speak (Furui, 1981).

One of the earliest approaches to recognising voices by technological means was the "voiceprint" (Kersta, 1962), which is simply the spectrogram of a given utterance. However, although subjective evaluation of voiceprints can provide success rates in excess of 98% for the identification of speakers (for small population sizes), it cannot be automated straightforwardly. Endres *et al.* (1971) describe several methods of extracting descriptive features from spectrograms.

Another way of describing the dynamic characteristics of a specific utterance is to construct a "time-varying" pattern of features that characterise short-term aspects of the speech sounds. Such features can include the pitch (Atal, 1972), LPC (reflection) coefficients (Furui, 1981), and cepstral coefficients (O'Shaughnessy, 1986). So that they can be matched effectively, dynamic patterns must be aligned such that the short-term features corresponding to each sound unit coincide. Atal (1972) linearly normalises each utterance to a fixed duration, but, because changes in speaking rate do not scale the duration of each phonetic segment equally, it is generally better to invoke some such technique as dynamic time-warping (DTW) (Furui, 1981; O'Shaughnessy, 1986). A practical system which employs DTW on the cepstral coefficients of a fixed phrase to verify the identity of a speaker is described by Birnbaum *et al.* (1986). Other techniques for matching the feature patterns include HMMs and neural networks, which are described in §3.6.1.

The characteristics of a speaker's voice include the long term (over several seconds) averages of features such as the pitch frequency, loudness, and spectral content (Markel *et al.*, 1977; Furui, 1981). Furthermore, the standard deviations of these speech characteristics over several seconds can indicate the "expressiveness' of the speaker (Markel *et al.*, 1977). The spectral content can be characterised by the LPC coefficients (Markel *et al.*, 1977) or by the spectrum (LTAS, see §3.4.2) itself (Doddington, 1985). Furui (1981) finds that recognition rates (for both identification and verification schemes) are similar whether "statistical" features or "dynamic" features are employed.

A technique of feature matching that takes into account both the long-term average features and the specific peculiarities of a person's pronunciation is to *vector quantise* all the (short-term) feature vectors of an utterance (Burton, 1987; Soong *et al.*, 1985). The resultant *code book* of cluster centroids (§3.6.1.3) represents the different types of sounds uttered by the person. Furthermore, the centroids differ for each person because of peculiarities in pronunciation of each phoneme. Hence, the code book obtained from each speaker can be thought of as a descriptor of the average characteristics of that person's voice and pronunciation.

3.6.3 Text-to-speech conversion

Text-to-speech conversion is the technical term for for automatic conversion of an arbitrary text into speech. It is invoked for various types of computer output, for book reading aids for blind people, and for voice storage (Klatt, 1987). In this section I overview what is required of text-to-speech conversion systems. Klatt's (1987) review comprehensively covers all other relevant details.

A text-to-speech system has two main functions. First, the system must analyse the text and determine its phonemic content (despite the vagaries of how sounds are spelt in different words). If the system is required to produce "natural sounding" speech, it must also analyse the syntax and semantics of the text. This is necessary so that it can determine the correct intonation and stress to apply to the speech. The details of such linguistic analysis are provided by Klatt (1987).

The second step in the text-to-speech conversion process is the transformation of the linguistic representation of the text into an acoustic representation. This requires both a synthesiser, which can produce the required quality of speech, and some method of converting the linguistic description of the text into control signals for the synthesiser.

Three types of synthesiser are commonly employed in text-to-speech systems. One of these is the *formant synthesiser*, which generates sounds from a description of them in terms of their formant frequencies and excitation (VUV and pitch). Synthesis of speech from its formant frequencies is discussed in §3.5.2.1. Straightforward rules can be invoked to convert the different phonemes, and combinations of phonemes, into their appropriate formant, pitch, and VUV control signals (Klatt, 1987).

Another type of synthesiser is directly modelled on the human speech production mechanism (§2.3.1). Such *articulatory synthesisers* are controlled by parameters such as tongue position, lip opening, nasality, and vocal cord tension. Although such synthesisers are attractive because of their potentially high voice quality (due to their close mimicry of human speech production), the reliable extraction of control parameters from text, and the operation of such synthesisers in real time (cf. Sondhi and Schroeter, 1987), must both be improved before these benefits can be fully realised.

A third approach to synthesising speech for text-to-speech conversion is simply to concatenate appropriate segments of real speech according to the linguistic information extracted from the text. The segments (which are generally compressed by an appropriate technique, such as LPC §3.5) may comprise single phonemes, diphones, or parts of syllables. Phoneme-based systems have the lowest storage requirements (about 40–60 segments of sound, corresponding to each of the different phonemes and allophones, are all that are required for an English language system), but produce speech that sounds very unnatural. This is because discontinuities often occur when different phonemes are concatenated. Also, such systems are unable to account for co-articulation (§2.1.4.2) effects between phonemes. Systems based on the concatenation of diphones (which are phoneme pairs) are able to account for co-articulation effects between these pairs. Typically, about 1000 diphones are required to produce reasonable sounding speech.

Apart from the ability of the linguistic analysis stage of a text-to-speech convertor to produce an accurate rendition of the content of a given text, the quality of speech from such a system depends on the sophistication of the speech synthesiser. Formant based synthesisers are currently the most flexible in producing natural sounding speech, although concatenation methods of synthesis are often able to generate speech of comparable quality. Improvements in the quality of speech generated by both these methods can be made by improving the "naturalness" of the voiced excitation source (cf. Holmes, 1973). For the concatenation schemes, improved methods of speech storage are required. These should provide the flexibility to alter the pitch and loudness levels

of the synthetic speech, while still producing speech of high quality.

3.6.4 Medical and therapuetic applications

In this section I discuss the use of speech analysis techniques in the diagnosis of certain voice disorders (§3.6.4.1), in providing aural and visual feedback during speech training(§3.6.4.2), and for assisting people with speech-related disabilities (§3.6.4.2).

3.6.4.1 Diagnosis of laryngeal dysfunction

The diagnosis of laryngeal disorders is commonly accomplished during clinical examination by such means as "laryngoscopes" and radiography (cf. Aronson, 1985, Chapter 4), or by subjective evaluation of patients' voice *quality* (Kasuya *et al.*, 1986). For example, laryngeal pathologies such as vocal nodules or vocal cord paralysis cause speech to have a "breathy" or "rough" quality (cf. Wendler *et al.*, 1986; Kasuya *et al.*, 1986). Indeed, experienced voice pathologists tend to be consistently reliable in diagnosing voice and laryngeal disorders by subjective evaluation of such voice qualities (cf. Wendler *et al.*, 1986). However, acoustic analysis of the speech signal is useful in that it allows a quantitative measurement of voice dysfunction. Several recent studies have examined the relationship between the subjective evaluation of voice quality by experienced clinicians and various quantitative features which characterise certain aspects of the speech signal (cf. Imaizumi, 1986; Fritzell *et al.*, 1986; Childers, 1990).

Several features have been invoked in attempts to characterise the "abnormal" character of breathy, rough, or otherwise dysfunctional voices (Kasuya *et al.*, 1986). The pitch often exhibits *perturbations* which contribute to the rough quality of a voice (cf. Lieberman, 1963; Maeda *et al.*, 1968). In addition, breathy or rough voices exhibit a high level of *vocal noise*, superimposed upon the vocalic sound. This noise arises because dysfunctional vocal cords often do not close completely (cf. Kasuya *et al.*, 1986). Various techniques have been developed for measuring the pitch perturbation during an utterance. Koike (1973, quoted in Kasuya *et al.*, 1986) defines a *perturbation quotient* as the ratio of the average variation of pitch between different pitch estimates and the average pitch period during the utterance. Using this measure, Kasuya *et al.* attempt to classify voices as "normal" or "pathologic" with only partial success (error rates of about 20%). Imaizumi (1986) employs a perturbation measure that describes the perturbations in the correlations between segments of speech in different pitch intervals. This measure shows a reasonable correlation with subjective evaluations of voice "roughness".

Kasuya *et al.* (1986) measure the amount of vocal noise which remains after the speech is filtered with an adaptive comb filter, thereby removing the harmonics of the fundamental pitch frequency. They again classify voices as normal or pathological, with error rates of about 20%. Banci *et al.* (1986) characterise the noise by measuring the energy in the LPC residue signal (as defined in §3.2.1). Their classification of pathological and normal voices is in error by 17% on average.

Other researchers have extracted the glottal waveform from the speech signal (§3.4) as an aid in characterising glottal function. Fritzell *et al.* (1986) employ an inverse filtering technique to obtain the glottal flow waveforms for several normal and pathological subjects. From the flow waveforms they measure the "leakage" airflow and hence obtain an estimate of the *glottal insufficiency*, or size of the glottal leakage opening. In addition, they compute the quotient of minimum to maximum airflow. Both of these measures correlate favourably with the subjective ratings of "breathiness" made by voice clinicians.

Several studies have examined the usefulness of the LTAS (see §3.4.2) as a descriptor of laryngeal function. Boves (1984, §5.3.5) shows that the LTAS characterises the average spectrum of the glottal excitation. Löfqvist (1986) characterises the LTAS by two measures: the ratio of energy below 1kHz to that in the band 1–5kHz; and the relative energy in the band 5–8kHz. The energy above 5kHz is supposed to indicate the level of noise in the speech signal while the low frequency ratio is a measure of the "slope" of the LTAS, which indicates the relative level of harmonic components in the glottal waveform (also see Sundberg and Gauffin, 1979). However, both Löfqvist and Wendler *et al.* (1986) find that their measures of LTAS do not distinguish between normal and pathological voices. The differences between normal speakers is greater than any differences due to pathologies. Despite this, the LTAS obtained from any particular speaker, throughout the course of treatment for a vocal disorder, tends to alter consistently with changes in the individual's disorder (Löfqvist, 1986). So it seems that the LTAS could be useful as a comparative measure of how someone's voice changes during therapy.

3.6.4.2 Therapy for people with voice disorders

In this section I briefly introduce the use of speech processing techniques as aids for people with speech difficulties. Further details can be found in the reviews by Damper (1982), Mangold (1988), and Childers (1990).

One application in this area is that of providing *bio-feedback* about certain speech characteristics in order to help people correct any deficiencies in their speech. The use of bio-feedback has been shown to improve the rate of learning for many types of learning tasks (cf. Rubow, 1984). In the context of speech therapy, an example is furnished by the learning of intonation by deaf people. Immediate visual feedback of the pitch and loudness of their speech sounds can assist them in attaining control over these parameters (cf. Levitt, 1973; Elder *et al.*, 1987; Bernstein *et al.*, 1988; and also see Scott and Caird, 1983). Another effective type of visual feedback of speech parameters is the real-time display of speech spectrograms (so-called "visible speech", cf. Kopp and Green, 1946) for helping people to learn to generate particular sounds correctly (Levitt, 1973).

Another application is the provision of *speech aids* to people who have speech or language communication difficulties. For people with speaking disabilities, the use of a portable and easily controllable synthetic voice can greatly improve their ability to communicate, and hence their quality of life (Thornett, 1989). Blind people experience many difficulties, some of the more serious of which are due to their inability to read things such as books, newspapers, computer displays, etc. Text-to-speech systems are beginning to make a real impact as "reading machines", especially since their performance and affordability continue to improve (cf. Klatt, 1987). Speech recognition devices are also useful for motor impaired people so that they can control their environment (cf. Fried-Oken, 1985; Mangold, 1988).

Aids for deaf people range from devices to transpose speech sounds into lower frequency bands, where their hearing is better (at least for those with only partial hearing loss), to techniques for translating speech into visual, tactile, or cochlear nerve patterns (Pickett, 1972; Levitt, 1973; Dormer and Phillips, 1987). Although such devices cannot as yet provide deaf people with sufficient levels of "aural" perception to understand everyday speech adequately enough for normal conversation, they can assist the deaf person's use of lip reading by providing additional information about the types of sounds being spoken (cf. Levitt, 1973; Kassling, 1989). Devices such as these are especially useful in situations when normal lip reading is not possible, such as when communicating by telephone. The future development of speaker independent

and large vocabulary speech recognition systems would facilitate the implementation of personal speech-to-text translation devices for deaf people (Damper, 1982).

Chapter 4

Shift-and-add processing of speech

Shift-and-add (SAA) is a method of blind deconvolution which is applicable when an ensemble of differently blurred signals are available (Bates and Cady, 1980). The technique arose as a method of removing the distorting effects of the atmosphere on astronomical images produced by large telescopes. §4.1 describes these astronomical origins of SAA. In §4.2 the source-filter model of speech production is re-formulated in a way that is more relevant to SAA processing. The technique of SAA, when applied to speech signals in the manner described in §4.2, is a method of extracting an estimate of the average glottal excitation "pulse". §4.3 presents some of the results obtained by applying SAA to synthetic and natural speech, and describes how the SAA signal can be refined to represent the glottal pulse shape more closely. SAA is a necessary step in the speech encoding technique introduced in Chapter 5. §5.4 of that chapter contains a discussion of the different ways in which the SAA signal can be interpreted. Some other applications where SAA may become useful are discussed in §8.2.1 of Chapter 8.

4.1 Astronomical background

Large telescopes are employed by observatories in order to gather more light for the observation of distant stars. In order to gather more light from faint objects, exposure times of many minutes or hours are common. However, the distortion introduced by the atmosphere (termed the *seeing problem*) means that the maximum resolution attainable by a large telescope (operating under average *seeing conditions* at a wavelength of about 500nm, which is near the middle of the visible spectrum) is about 1 arcsec, or no better than that possessed by one with a 10cm diameter (Bates, 1982; Roddier, 1988). By comparison, the theoretical resolution (diffraction limit) for a 4m diameter telescope is about 0.025 arcsec for light of wavelength equal to 500nm (Davey, 1989, §4.1). The distortion is due to continual and random phase fluctuations introduced by turbulence in the atmosphere (cf. Strohbehn, 1968). The image of an *ideally unresolvable* object (i.e. unresolvable even under ideal seeing conditions) that is formed by a long exposure is termed the *seeing disk*. It is worth noting that very few stars are ideally resolvable even in the largest telescopes.

During short exposures (typically of 10ms duration), the atmosphere is effectively frozen (Davey, 1989, §4.2). The resulting images have a speckly appearance and so are called *speckle images* (Bates, 1982). A speckle image $s_m(\mathrm{x})$ can be considered as a convolution between the *ideal image*, or *object* $f(\mathrm{x})$ and a (random) *blurring function*

$h_m(\mathbf{x})$ which characterises the atmosphere and telescope (Davey, 1989, pp53–55):

$$s_m(\mathbf{x}) = f(\mathbf{x}) \odot h_m(\mathbf{x}) + c_m(\mathbf{x}). \tag{4.1}$$

The quantity $c_m(\mathbf{x})$, which is termed the contamination, characterises all deviations from the convolutional model. The subscript m indicates that $s_m(\mathbf{x})$ is one of an ensemble of M such speckle images, recorded sequentially so that the atmospheric blurring function $h_m(\mathbf{x})$ is essentially statistically independent between each pair of measurements. The variable $\mathbf{x}$ is in general a two-dimensional vector. Note also that all quantities appearing in (4.1) are *positive* (i.e. non-negative real) because optical astronomical objects are spatially incoherent.

4.1.1 Astronomical shift-and-add

Shift-and-add (SAA) (Bates, 1976; Bates and Cady, 1980) is a method of ensemble blind deconvolution (Davey, 1989, Chapter 3). SAA processing of the speckle images described above involves (a) locating the brightest point $\mathbf{x}_m$ (termed the *SAA reference*) in the m^{th} speckle image, (b) shifting each $s_m(\mathbf{x})$ so that the brightest point is at the origin, and (c) averaging together the ensemble of M speckle images (Davey, 1989, §4.8.1). The resulting SAA image $f_{sa}(\mathbf{x})$ can be expressed as

$$\begin{aligned} f_{sa}(\mathbf{x}) &= \langle s_m(\mathbf{x}+\mathbf{x}_m)\rangle_m \\ &= f(\mathbf{x}) \odot \langle h_m(\mathbf{x}+\mathbf{x}_m)\rangle_m + \tilde{c}(\mathbf{x}) \\ &= f(\mathbf{x}) \odot h_{sa}(\mathbf{x}) + \tilde{c}(\mathbf{x}) \end{aligned} \tag{4.2}$$

where $\tilde{c}(\mathbf{x})$ is the contamination and $h_{sa}(\mathbf{x})$ is the SAA blurring function. M should be large enough to ensure that spatial variations exhibited by $f_{sa}(\mathbf{x})$ are smaller than those in $f(\mathbf{x})$. Fig.4.1 shows an example of SAA applied to computer-generated speckle images of an (artificial) object. This illustrates how the components of the star are recovered from the speckle images, but super-imposed upon a *fog*, which is similar in form to the seeing disk mentioned in the first paragraph of this section.

The fog can be removed from the SAA image of the object by generating a SAA image from a single ideally unresolvable star (under seeing conditions statistically similar to those pertaining when the object was observed) and deconvolving this estimate of the fog from the SAA image (cf. Davey, 1989, pp66–67). Fig.4.2 shows the result of this type of "defogging' applied to the SAA image shown in Fig.4.1*d*.

4.1.2 Ghosts in Shift-and-add

The basic SAA defined in (4.2) is successful in recovering the true image when the object contains a single, unresolvable, point which is much brighter than all the other points in the image (Davey, 1989, §4.8.2). This is because each speckle image $s_m(\mathbf{x})$, defined the convolution (4.1), can be considered as many copies of $f(\mathbf{x})$, with $h_m(\mathbf{x})$ specifying the weighting applied to each copy. If the brightest point in $f(\mathbf{x})$ *dominates* all the other points of the object, the brightest point in $s_m(\mathbf{x})$ (the SAA reference) then corresponds to that same point in the brightest copy of the object. Hence the copies of the object that are centred by the "shift" of SAA from each of the speckle images add constructively and $f(\mathbf{x})$ is recovered faithfully.

For objects that are not dominated by a single bright point, the SAA reference in each speckle image may not always correspond to the brightest point in the brightest copy of the object. When this occurs, the brightest copy of the object is shifted to the "wrong" place relative to the origin and $f(\mathbf{x})$ is not faithfully recovered. For example,

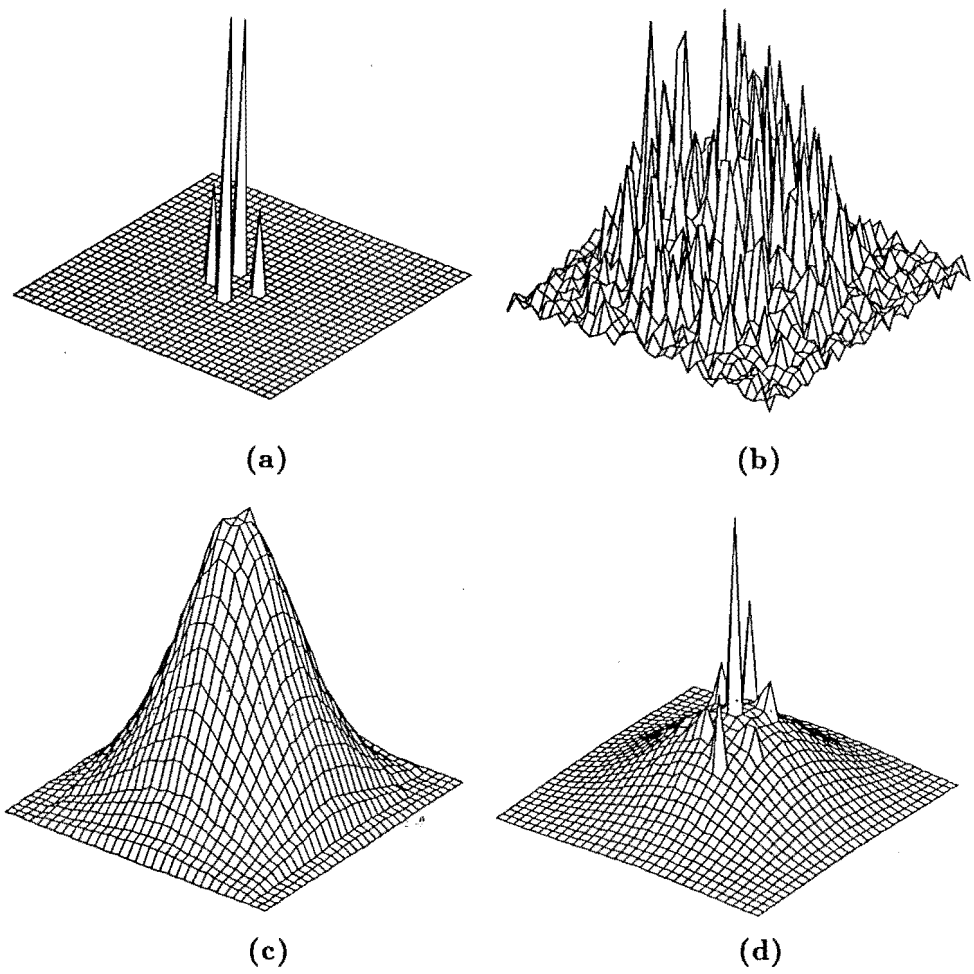


Figure 4.1. SAA of computer-generated speckle images of an object having several distinct peaks. **a:** The ideal image $f(\mathbf{x})$. **b:** A typical computer-generated speckle image. **c:** The seeing disk formed by adding (without shifting) 1024 speckle images such as the one shown in *b*. **d:** $f_{sa}(\mathbf{x})$ of 1024 speckle images similar to that shown in *b* (these images copied with permission from Bates and Davey, 1987).

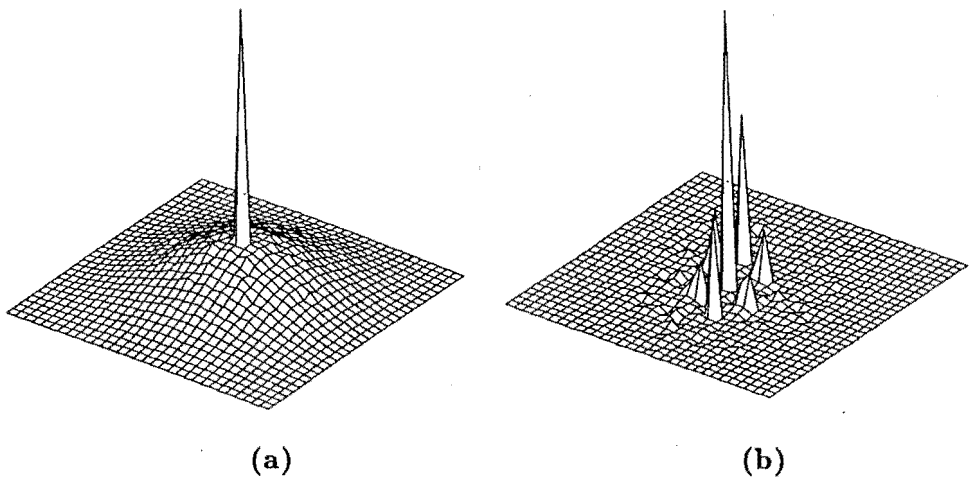


Figure 4.2. **a:** SAA of an unresolvable star. **b:** The SAA image shown in Fig.4.1*d* "defogged" by deconvolving out the SAA shown in *a* (after Bates and Davey (1987), with thanks).

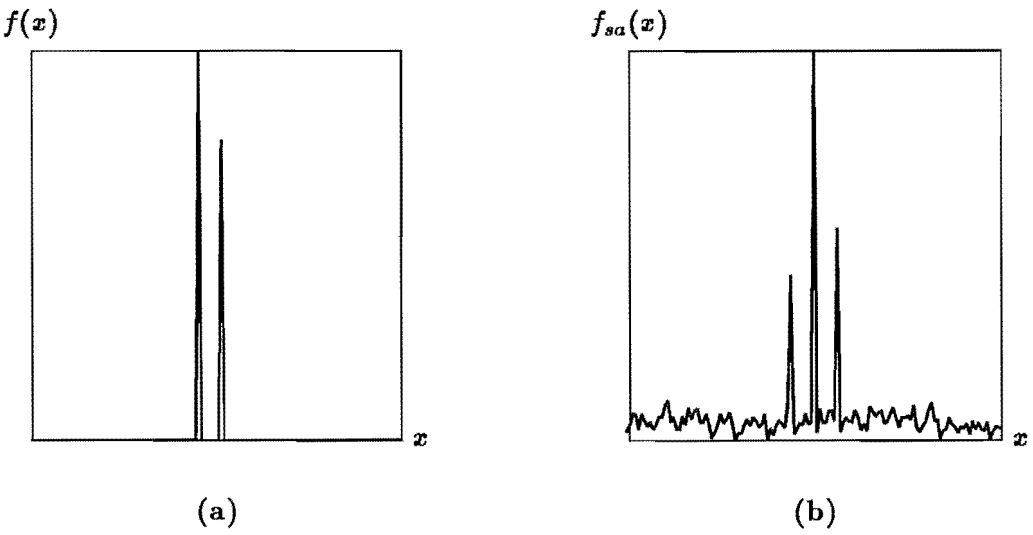


Figure 4.3. Ghosts in SAA images. **a:** Object, containing two spikes of unequal, but comparable, magnitudes. **b:** SAA image, obtained from an ensemble of computer generated speckle images, each a different distortion of the object shown in *a*. The ghost is the third spike that appears on the left of the SAA image.

Fig.4.3*a* illustrates an image consisting of two non-zero "spikes" (such as might represent a binary star). In the ensemble of speckle images, the SAA reference sometimes refers to the left-most point of the brightest copy of the object, and sometimes to the right-most point. These events occur in the same proportion as the relative magnitudes of the spikes. In consequence of this, the SAA image, Fig.4.3*b*, contains three spikes instead of two. The third spike (termed a *ghost*) results from the constructive superposition of the copies of $f(\mathbf{x})$ that were centred on the wrong peak (Bates and Cady, 1980).

The amplitudes of the ghosts depends on the form of the object and the relative dominance of the major peak (Hunt *et al.*, 1983, §§4,5). Because of the object-dependence of the ghosting, it is convenient to separate the SAA blurring function into an object-dependent component $h_{sa_o}(\mathbf{x})$ and a seeing-dependent component $h_{sa_s}(\mathbf{x})$ (Bates and Davey, 1987). $h_{sa}(\mathbf{x})$ is then described by

$$h_{sa}(\mathbf{x}) \approx h_{sa_o}(\mathbf{x}) \odot h_{sa_s}(\mathbf{x}). \tag{4.3}$$

The fog mentioned in the previous paragraph is seen to correspond to $h_{sa_s}(\mathbf{x})$, while the ghosting corresponds to $h_{sa_o}(\mathbf{x})$. Various techniques have been proposed for removing the ghosting and improving the faithfulness of the SAA images (Bates and Davey, 1987; Davey *et al.*, 1989; Davey, 1989, §4.8.3). However, each of these extensions adds significantly to the amount of computer processing required.

4.1.3 Other applications of shift-and-add

SAA has also been applied to ultrasonic imaging (Bates and Robinson, 1981). Ultrasonic images are obtained by measuring the acoustic waves scattered from an object when it is subjected to an incident field of ultrasonic radiation. The true image $f(\mathbf{x})$ of the equivalent object is the distribution of the complex scattering amplitude throughout the insonified part of the acoustic medium. The images have a speckly appearance, which is different for images recorded in different frequency bands (Bates and Robinson, 1981). Thus SAA can be performed on the ensemble of speckle images so produced to

recover a representation of the true image. Because the speckle images are complex-valued, the SAA process is modified by multiplying each speckle image by the phase of the brightest point:

$$f_{csa}(\mathbf{x}) = \langle s_m(\mathbf{x} - \mathbf{x}_m) e^{i\phi(\mathbf{x}_m)} \rangle_m, \tag{4.4}$$

where $\phi(\mathbf{x}_m)$ is the phase of the brightest point. Because the object and the blurring functions are spatially *coherent* (as manifested by their possession of phases as well as magnitude), an ultrasonic $f_{csa}(\mathbf{x})$ tends to be fog-free (the fog is self-cancelling by destructive interference).

4.2 SAA for speech signals

The reasoning which led to the development of SAA for astronomical speckle imaging can also be applied (with a suitable modification of terminology) to speech signals. This section develops that reasoning, and in doing so presents the mathematical basis for performing SAA analysis on speech signals (§4.2.1). Some of the characteristics of speech signals that influence SAA processing of speech are discussed in §4.2.2. The SAA algorithm for speech signals is presented in §4.2.3. Finally, details of the implementation are described further and in depth in §4.2.4.

A speech signal (initially I limit the discussion to voiced speech) can be considered as a convolution between a series of glottal pulses and a time-varying vocal tract filter (§2.3.1.4). By likening each pitch interval of the speech signal (which contains a "blurred" copy of the "archetypal" glottal pulse) to one of the speckle images observed in speckle astronomy, the technique of SAA can be adapted to extract the "glottal pulse" from its many differently-blurred manifestations that occur during a typical speech utterance. In short, the technique of SAA, as described in §4.1, is applied to speech by making the following metaphorical relations between astronomical and vocal quantities: the true astronomical object with the archetypal glottal pulse; the randomly varying atmospheric distortions with the random (!) distortions induced by the vocal tract; and each speckle image with each pitch-period length segment of speech. Note that the variations in the shapes of individual glottal pulses are considered as distortions of the archetypal pulse shape. These distortions are assumed to be less than the distortions introduced by the vocal tract (see §4.2.2.1). §4.2.1 expands this model of the speech signal, in order to elucidate the characteristics of SAA processing on speech.

4.2.1 Mathematical description

A speech utterance consists of many, say M, segments $s_m(t)$, each containing a single period of voiced speech, together with other segments, suitably juxtaposed, containing unvoiced speech which I shall conveniently ignore until §4.2.4.5. The (voiced) speech signal $s(t)$ is therefore described by

$$s(t) = \sum_{m=1}^{M} s_m(t - T_m) \tag{4.5}$$

where T_m identifies the instant at which the magnitude of $s_m(t)$ is greatest. Taking the source-filter approach, each speech segment is further described by a convolution between the excitation pulse $g_m(t)$ during the m^{th} segment and the causal response of the vocal tract filter $v_m(t)$:

$$s_m(t - T_m) = g_m(t - T_{em}) \odot v_m(t - T_m + T_{em}) \tag{4.6}$$

where T_{em}, the instant at which $|g_m(t)|$ is greatest, is introduced because, due to the "blurring" effects of $v_m(t)$, it is in general different from T_m.

It is convenient to separate $g_m(t)$ and $v_m(t)$ into *invariant* and *variant* components. The invariant component is defined to be the average of all the signals $g_m(t)$ or $v_m(t)$ respectively. The variant component of each signal can be further divided into *convolutional* and *additive* parts. For instance, the excitation pulse $g_m(t)$ is described by

$$g_m(t - T_{em}) = g^i(t - T_{em}) \odot g^v_m(t) + g^c_m(t) \tag{4.7}$$

where $g^i(t)$ is the invariant component, $g^v_m(t)$ is the convolutional part, and $g^c_m(t)$ is the additive part of the variable component of the m^{th} pulse $g_m(t)$. The invariant excitation is defined to be

$$g^i(t) = \langle g_m(t + T_{em}) \rangle_m, \quad -\tau_- < t < \tau_+, \tag{4.8}$$

where $(\tau_- + \tau_+)$ is the effective duration of $g^i(t)$.

The convolutional component $g^v_m(t)$ can be considered as modelling the linear deviations of $g_m(t)$ from $g^i(t)$, while the additive component $g^c_m(t)$ models any non-linear deviations that cannot be accommodated by the convolution. Because of the general inconsistency of convolution (§1.2.5.3), there are an infinite number of signals $g^v_m(t)$ and $g^c_m(t)$ that satisfy (4.7). Further constraints must be introduced to ensure that they can be unique (see §4.2.2 and §5.4).

The vocal tract is represented in the source-filter model as a time-varying filter. However, for short intervals it may be considered as time-invariant (§2.3.1.3). Hence it can be likened to the atmospheric blurring that occurs in optical astronomical speckle images. Because the vocal tract does not in general vary in a completely unbiased manner (cf. §4.2.2.1), it is appropriate to separate it into variant and invariant parts. The vocal tract response $v_m(t)$ for the m^{th} speech segment is described by

$$v_m(t) = v^i(t) \odot v^v_m(t) + v^c_m(t), \tag{4.9}$$

with $v^v_m(t)$ and $v^c_m(t)$ being the convolutional and additive deviations, respectively, from the invariant part $v^i(t)$. This invariant part is the "average" vocal tract response, which depends upon the utterance being spoken, the physiology of the speaker's vocal tract, and nuances of the speaker's articulation (§2.1.4.1). It is useful, for the purposes of SAA, to define the invariant component $v^i(t)$ as

$$v^i(t) = \langle v_m(t + T_m - T_{em}) \rangle_m. \tag{4.10}$$

The shift-and-add signal $s_{sa}(t)$ is defined by

$$s_{sa}(t) = \langle s_m(t + T_m) \rangle_m, \quad -\tau^a_- < t < \tau^a_+, \tag{4.11}$$

where $(\tau^a_- + \tau^a_+)$ is the duration of the average pitch interval for the utterance. Invoking the definitions (4.7) and (4.9) allows (4.11) to be expressed as

$$\begin{aligned} s_{sa}(t) &= \langle g_m(t + T_{em}) \odot v_m(t + T_m - T_{em}) \rangle_m & (4.12) \\ &\approx g^i(t) \odot v^i(t). & (4.13) \end{aligned}$$

It is useful to call $s_{sa}(t)$ the *invariant* speech component, rewriting it for notational consistency as $s^i(t)$. Taking (4.13) to be exact rather than merely approximate (see §4.2.2.2) allows the speech segment $s_m(t)$ to be expressed as

$$s_m(t - T_m) = s^i(t - T_m) \odot s^v_m(t) + c_m(t), \quad T_m - \tau_{m-} < t < T_m + \tau_{m+}, \tag{4.14}$$

where

$$s_m^v(t) = g_m^v(t) \odot v_m^v(t), \tag{4.15}$$

is termed the *generalised speech filter* component and

$$s_m^c(t) = v_m^c(t) \odot g^i(t) \odot g_m^v(t) + g_m^c(t) \odot [v_m^c(t) + v^i(t) \odot v_m^v(t)] \tag{4.16}$$

is called the *contamination* term (Brieseman *et al.*, 1987). The quantities τ_{m-} and τ_{m+} define the extents of the m^{th} pitch interval, before and after, respectively, the instant T_m of maximum speech magnitude. The limits τ_{m-} and τ_{m+} mean that $s_m(t)$ is effectively a truncated version of the convolution between $s^i(t)$ and $s_m^v(t)$. Consequently, the contamination $s_m^c(t)$ also includes the parts of $s^i(t) \odot s_{m-1}^v(t)$ that overlap into the m^{th} segment.

It is important to recognise that there are many possible ways of defining $s_m^v(t)$ and $s_m^c(t)$ so that they satisfy (4.14). Further constraints are necessary for $s_m^v(t)$ and $s_m^c(t)$ to be unique. As regards SAA processing, the relative contributions of $s_m^v(t)$ and $s_m^c(t)$ to $s_m(t)$ are unimportant. For the kind of processing described in Chapter 5, however, $s_m^v(t)$ must dominate $s_m^c(t)$ in some useful sense. The constraints that the characteristics of speech signals place on the relative contributions of $s_m^v(t)$ and $s_m^c(t)$ are discussed in detail in §5.4.1.

4.2.2 Speech characteristics relevant to SAA processing

Application of SAA processing to speech signals, in the manner described in §4.2.1, requires several assumptions to be made about the nature of speech sounds and how they are to be mathematically represented. §4.2.2.1 examines the validity of representing speech signals in terms of variant and invariant components, while §4.2.2.2 discusses the characteristics of the glottal pulse and vocal tract response that affect the success of SAA on speech. There is further discussion of these topics in §5.4.1, with emphasis on the implications of SAA combined with CLEAN processing for speech.

4.2.2.1 Variability and invariance in speech

In the source-filter model of speech espoused in §4.2.1, each segment of speech $s_m(t)$ contains a glottal source component $g_m(t)$ and a vocal tract filter component $v_m(t)$. Within any one of these segments, which are typically of 5ms to 10ms in duration, both $g_m(t)$ and $v_m(t)$ can be considered as time-invariant impulse response functions. However, in a typical utterance, they both vary considerably between different segments.

Variability in the vocal tract impulse response arises from the different articulatory configurations employed to pronounce different phonemes, whereas the variability in glottal excitation results mainly from changes in pitch, vocal intensity, and vocal stress (§2.1.4). In a typical utterance, then, the variability in the speech signal due to changes in the articulatory configuration is greater (and more linguistically relevant) than that due to changes in the glottal excitation. Consequently, as a first approximation, one can view the glottal pulse $g_m(t)$ as invariant and the vocal tract $v_m(t)$ as variable with $\langle v_m(t) \rangle_m$ negligible. The invariant component $s^i(t)$ of an utterance then reduces to the invariant glottal pulse $g^i(t)$ while the variant component $s_m^v(t)$ becomes the vocal tract response $v_m^v(t)$. Further refinements of this approximation are considered in §4.3 and §5.4.1.

The variation of $g_m(t)$ and $v_m(t)$ between different segments is not "random", in the sense that tossing a coin is random, because it is related to the linguistic and para-linguistic content of the utterance (cf. §2.1.3.1). Thus one expects closely spaced segments to be more highly correlated than those that are further apart. However,

the variation can be called random in the sense that, assuming the utterance to be *phonetically balanced* (i.e. it is not weighted in favour of any particular type of sound), the difference between any randomly selected pair of segments is effectively "random". An utterance which is phonetically balanced in this way is termed "unbiased" in this thesis.

4.2.2.2 Ghosting in SAA processing of speech

In this section I discuss the characteristics of speech that affect the "peakiness" of the invariant speech component. As mentioned in §4.1.2, SAA is most successful when the object contains a dominant peak. The peakiness of the invariant component is determined mainly by the form of the glottal pulse, as pointed out in §4.2.2.1. The characteristics of the "blurring" induced by the vocal tract also affect the ability of SAA to extract a faithful replica of the true pulse, since the form of blurring influences how accurately the peak of the (largest copy of the) glottal pulse can be identified.

It is convenient here to employ the traditional source-filter terminology, with the assumptions made in §4.2.2.1 that $s^i(t) = g^i(t)$ and $s^v_m(t) = v^v_m(t)$. Each segment of a speech signal, viewed as a convolution between a glottal pulse and the vocal tract filter, can be regarded as many copies of the glottal pulse, each shifted and weighted differently. Although the convolutional blurring "smooths out" the peakiness of the glottal pulses, there is a certain probability that the position of greatest magnitude in the speech segment $s_m(t)$ corresponds to the peak of the largest copy of the glottal pulse $g^i(t)$. The success of SAA is predicated on this probability being appreciable enough that the instances when the largest copy of $g^i(t)$ is correctly identified dominate in the SAA average. The copies of $g^i(t)$ which do not have their peaks aligned with the largest magnitude in $s_m(t)$ tend to cancel out if they are randomly distributed. Conversely, any *systematic* placement of non-centred copies of $g^i(t)$ gives rise to ghosting, in the way described in §4.1.2.

As mentioned in §4.1.2, astronomical SAA (by itself) is most successful at recovering true images that contain a dominant, unresolvable peak. However, by incorporating additional processing steps into the recovery procedure (see Ayers and Dainty, 1988 and Davey *et al.*, 1989 for details), an individual blurred multi-dimensional image can almost always be blindly deconvolved uniquely into its constituent components. On the other hand, unique blind deconvolution of a one-dimensional image, or signal, is in general impossible (Lane and Bates, 1987). So, SAA can only really restore a blurred one-dimensional signal if it contains a dominant, unresolvable peak. In the context of speech signals, I use the term "impulsive" instead of "unresolvable" to refer to a peak that is effectively of one sample duration (i.e. it has a flat spectrum from d.c. to half the sampling frequency). The peak is dominant if its magnitude is much greater than any of the other samples of which the signal is comprised.

The acoustic pressure field outside the head is related to the acoustic waveform at the lips by the radiation characteristics of the lips, which can be approximated by first order differentiation (§3.2.1). This means that a recorded waveform is actually almost the differential of the acoustic waveform inside the vocal tract. Hence, the glottal pulse train of the recorded speech signal is more "peaky" than the "real" glottal flow waveform (see §3.4.1). This peakiness accords well with the requirements of SAA (§4.1).

The glottal pulse, while it is peaky due to the differentiation effect of lip radiation, does not possess a dominant impulsive peak. Nevertheless, by differentiating the speech signal, the *effective glottal pulse*, which is the second derivative of the glottal flow waveform, can be considered to be approximately impulsive (cf. §3.4.1; Anantha-padmanabha and Yegnanarayana, 1979), In §4.2.4.7 I present results of SAA performed

on differentiated and undifferentiated speech. These indicate that differentiating the speech signal does seem to "improve" the SAA signal in that it appears to correspond more closely with glottal waveforms obtained by other techniques.

As well as being affected by the shape of the glottal pulse, the severity of ghosting in SAA is affected by the characteristics of the vocal tract impulse response. The vocal tract filter is comparatively more narrow-band than the point spread function of the atmosphere which manifests itself in astronomical SAA (§4.1). The point spread function of the atmosphere has a relatively flat magnitude response, its blurring arising from phase distortion of the wave front as it passes through the atmosphere (cf. Bates, 1982). Because of this, the point spread function can be considered to be a number of impulses, randomly scattered over roughly the same extent as the seeing disk (Bates, 1982). By contrast, the magnitude response of the vocal tract filter consists of a number of relatively narrow resonance peaks. The width and centre-frequencies of these resonances vary according to the particular sound being uttered (cf. Fant, 1960). The effect of the relative narrowness of the vocal tract filter is to effectively broaden the peak of the glottal pulse. Hence, the position of maximum speech amplitude is less likely to correspond to the actual peak of the largest copy of the glottal pulse than in the astronomical case. In practice, the narrow-band nature of the vocal tract filter does not appear to adversly affect the performance of SAA. This is illustrated in §4.3.2, which presents results of performing SAA on speech-like sounds constructed from synthetic "vocal tracts" of various bandwidths.

By contrast with the atmospheric blurring, the vocal tract filter is imperfectly random. This can be most simply dealt with by incorporating any bias in the variation of the vocal tract response into the SAA signal, by means of the definition of $s^i(t)$ introduced in §4.2.1.

The "ghosting" effects described in the previous paragraphs mean that the resulting SAA signal $s_{sa}(t)$, instead of equalling $s^i(t)$ as implied by (4.13), is equal to

$$s_{sa}(t) = s^i(t) \odot h_{sa_g}(t) \tag{4.17}$$

where $h_{sa_g}(t)$ is the SAA ghosting function, its "severity" being determined by the magnitude of the effects discussed above (cf. §4.1.2).

Note that SAA for speech does not exhibit the fog which characterises astronomical SAA. The latter fog arises because astronomical images are positive, implying that the SAA image can only increase as more speckle images are added. By contrast, SAA for speech is "coherent" (cf. Bates and Robinson, 1981; see also (4.4) in this chapter) so that "noise" and non-centred copies of the glottal pulse tend to average to zero by destructive interference.

4.2.3 SAA algorithm for speech

As §4.2.1 indicates, SAA is performed on segments $s_m(t)$, each containing a single pitch interval of voiced speech extracted from a complete utterance $s(t)$. One of the attractive aspects of SAA is that these segments do not need to be explicitly identified before applying the SAA algorithm. This is because the segments can be identified during the SAA algorithm itself, as described in the next paragraph. Hence there is no need for accurate pitch or VUV analysis.

It is necessary to modify the SAA algorithm described in §4.1 so that it can be applied with good effect to speech signals. This is because the individual "speckle images" actually occur as sequential segments of a single speech signal. Hence the two steps, of segmenting the speech signal, and shifting each segment so its largest peak is at the origin, must be combined into one algorithm. The first step in the algorithm is

Parameter	Default value (units)	Comment
τ_s	12.8 (ms)	Segment duration
τ_p	10.0 (ms)	Segment spacing
η_{sa}	1.0 (% of peak)	Silence threshold value
η_{uv}	15.0 (% of peak)	"Unvoiced" threshold value

Table 4.1. Suggested default values for the parameters in the SAA algorithm described in §4.2.3.

to set $\tau_1 = \tau_p$, where τ_p is an estimate of the average pitch within the utterance (see §4.2.4.1 for suitable values for τ_p) Thereafter, for the m^{th} segment,

1. The maximum magnitude peak in $\tilde{s}_m(t)$, which is the segment of $s(t)$ delineated by the limits $\tau_m < t < \tau_m + \tau_s$, is denoted by T_m:
$$T_m = \text{argmax} \left|s(t)\right| , \tau_m < t < \tau_m + \tau_s, \tag{4.18}$$
where τ_s is the segment duration (see §4.2.4.1 for suitable values for τ_s).

2. The m^{th} *SAA segment* $\hat{s}_m(t)$ defined by
$$\hat{s}_m(t) = s(t + T_m - \tau_s/2) \ , \ 0 < t < \tau_s \tag{4.19}$$
is extracted from $s(t)$.

3. The start of the $(m+1)^{\text{th}}$ *search segment* $\tilde{s}_{m+1}(t)$ is located at
$$\tau_{m+1} = T_m - \tau_s/2 + \tau_p. \tag{4.20}$$

4. Steps 1,2 and 3 are repeated until $\tau_{m+1} + \tau_s$ exceeds the duration of the utterance.

5. The SAA signal is constructed by normalising and averaging together each of the $\hat{s}_m(t)$ for which $|\hat{s}_m(0)| > \eta_{sa}$, where η_{sa} is a threshold set to exclude segments that contain inter-word "silence" (also see §4.2.4.5):
$$s_{sa}(t) = \langle \hat{s}_m(t)/\hat{s}_m(0) \rangle_m, \ -\tau_s/2 < t < \tau_s/2. \tag{4.21}$$

§4.2.4, in addition to presenting detailed results arising from the application of this algorithm to various speech signals, describes the effects of altering the various parameters of the algorithm. Typical values for these parameters are listed in Table 4.1.

4.2.4 Implementation considerations

In this section I describe the results obtained by applying SAA processing to speech under various conditions. These conditions include both the various parameters of the algorithm presented in §4.2.3, and the characteristics of any particular utterance that the algorithm may be called upon to process. §4.2.4.1 describes the effect on the resulting SAA signal of varying the duration and spacing parameters, while §4.2.4.2 presents results illustrating the effect of changing the duration and content of an utterance. In §4.2.4.3 I discuss the ill-effects that result from phase distortion of the speech signal, and describe a method by which they can be ameliorated. §4.2.4.4 describes the effect that additive noise has on the SAA signal obtained from an utterance. The question of whether to normalise each segment of speech before averaging is considered in §4.2.4.6, while §4.2.4.5 discusses ways in which the unvoiced sections of an utterance can be removed or processed. Finally, §4.2.4.7 describes the apparent improvement in SAA performance that occurs if the speech signal is pre-emphasised by first-order differentiation before SAA is performed.

4.2.4.1 Choice of segment duration and spacing

The parameters τ_s and τ_p which represent, respectively, the duration of each of the speech segments comprising a particular utterance and the estimated interval between successive segments, are said in §4.2.3 to be equal to the available estimate of the average pitch period during the utterance. It is, however, computationally highly convenient for the values of τ_p and τ_s to be fixed for all utterances. In this section I present results which illustrate the changes in $s_{sa}(t)$ that arise from varying τ_p and τ_s. Since these results indicate that $s_{sa}(t)$ is relatively insensitive to changes in τ_p and τ_s, fixing their values for all utterances seems to be justified.

The segment duration τ_s must be no less than the average pitch period for the utterance being processed, in order to comply with (4.21). Fig.4.4 shows that the shapes of the central features of $s_{sa}(t)$ are little dependent upon the value of τ_s. In fact, the only noticeable effect of changing τ_s is to change the duration of $s_{sa}(t)$. It is convenient, therefore, to choose τ_s to be somewhat greater than the average pitch period of all the utterances which are to be processed. The duration of each $s_{sa}(t)$ can then be trimmed according to the requirements of any further processing that is to be performed on it (see, for example, §5.2.5.1).

Fig.4.5 shows several SAA signals obtained with different values for τ_p. These are all similar, except that the number of segments M from which each $s_{sa}(t)$ is formed is smaller in the cases when τ_p is larger. The details are discussed in the next few paragraphs. §4.2.4.2 presents more details on what effects different values of M have on the form of $s_{sa}(t)$.

The "segmentation" of $s(t)$ into its constituent segments $s_m(t)$ occurs as part of the SAA algorithm described in §4.2.3. Because the consequences of this segmentation are largely governed by the values assigned to τ_p and τ_s, it is appropriate to examine the types of *segmentation errors* that can occur for various extreme values of τ_p and τ_s.

One type of segmentation error occurs when any particular pitch period p_m within an utterance is greater than $\tau_p + \tau_s/2$. As illustrated in Fig.4.6*a*, this means that the search segment $\tilde{s}_{m+1}(t)$ does not encompass the position of the largest peak in the $(m+1)^{\text{th}}$ pitch interval. Hence the segment $\hat{s}_{m+1}(t)$, that is added to the SAA average, is not centred on the "true" instant of glottal excitation. However, for typical values of τ_s and τ_p (see Table 4.1), this type of error only occurs if the pitch frequency falls below 57Hz, which only applies to a small number of pitch intervals in utterances by a few male speakers.

Another type of error that can occur is illustrated in Fig.4.6*b*, which shows an instance of a maximum of $|s(t)|$ occurring slightly outside the segment $\tilde{s}_m(t)$. In this example no sample within $\tilde{s}_m(t)$ has a magnitude greater than that of the sample at the end-point of $\tilde{s}_m(t)$ nearest to the above-mentioned maximum. Because this sample is not actually a true peak in the signal, using it as the SAA reference causes ghosting of the SAA signal. This type of error can cause a small "bump" to appear on the side of the main peak of $s_{sa}(t)$.

Fig.4.6*c* shows another type of segmentation error that can occur if $p_m < (\tau_p - \tau_s/2)$, implying that the algorithm skips a complete pitch interval. This is of less concern than the errors mentioned in the previous two paragraphs, since it simply means that a longer utterance is required to generate a $s_{sa}(t)$ from a given number M of segments.

Because the errors mentioned above occur infrequently for typical utterances, their effects can effectively be ignored in practical applications of SAA (as is attested by the consistency of the SAA signals shown in Figs.4.4 and 4.5). This is because any "glitches" caused by the errors are "averaged out" by the averaging inherent in the

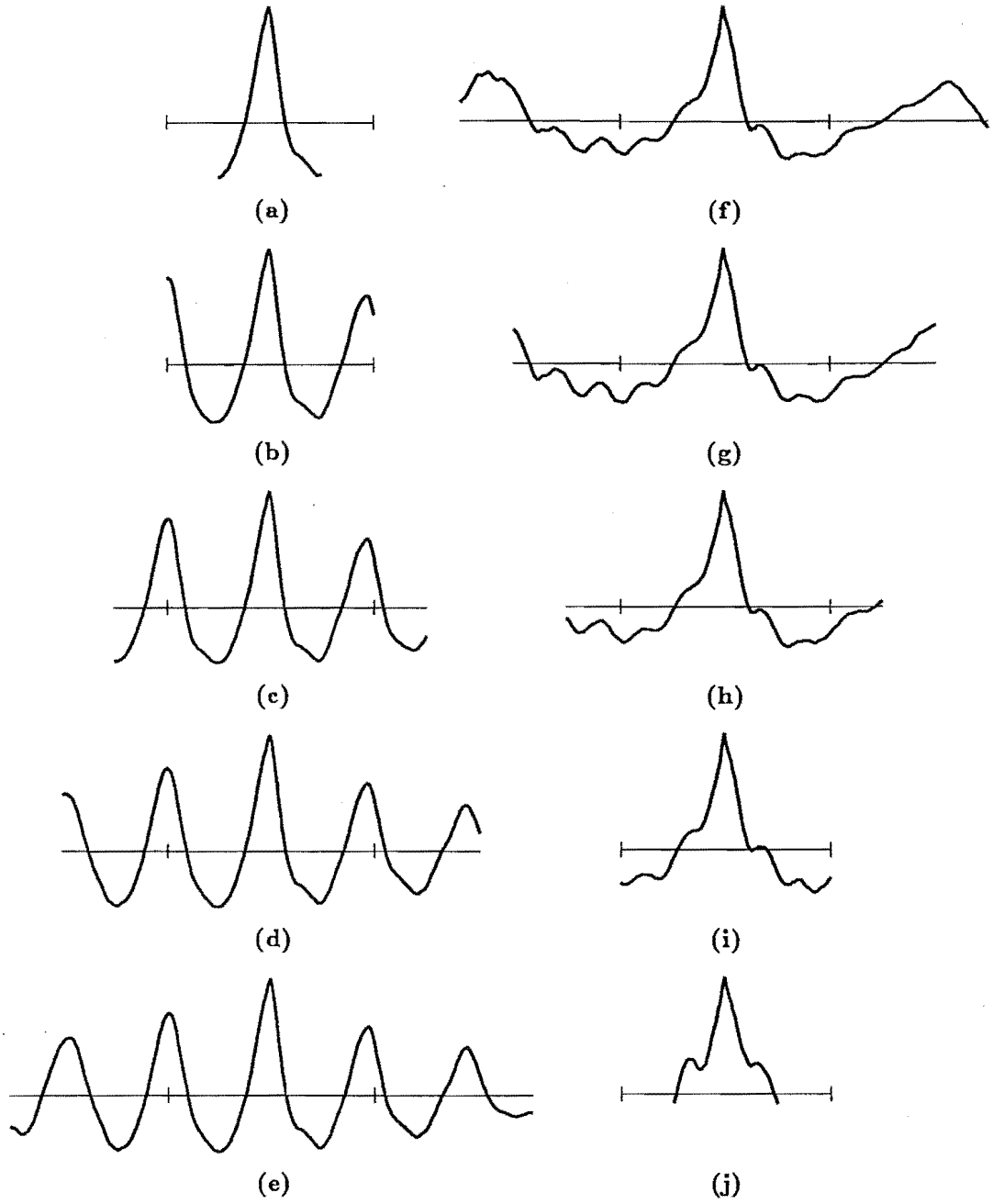


Figure 4.4. SAA signals obtained when the segment duration τ_s is varied, from 5ms to 30ms. In each case $\tau_p = 3/4\tau_s$. Utterance TF-RAIN1 (see Table 1.2): τ_s = **a:** 5ms, **b:** 10ms, **c:** 15ms, **d:** 20ms, and **e:** 30ms. Utterance WM-RAIN1: τ_s = **f:** 30ms, **g:** 20ms, **h:** 15ms, **i:** 10ms, and **j:** 5ms. All the SAA signals shown here, and in the remainder of this thesis, are plotted to the same time scale. The horizontal line on each signal represents the dc level of that signal, and the small vertical "tick marks" on this line indicate plus and minus 5ms (assuming that the SAA signals are centred on the time origin). The amplitude of each signal is arbitrarily normalised so that its peak value is unity.

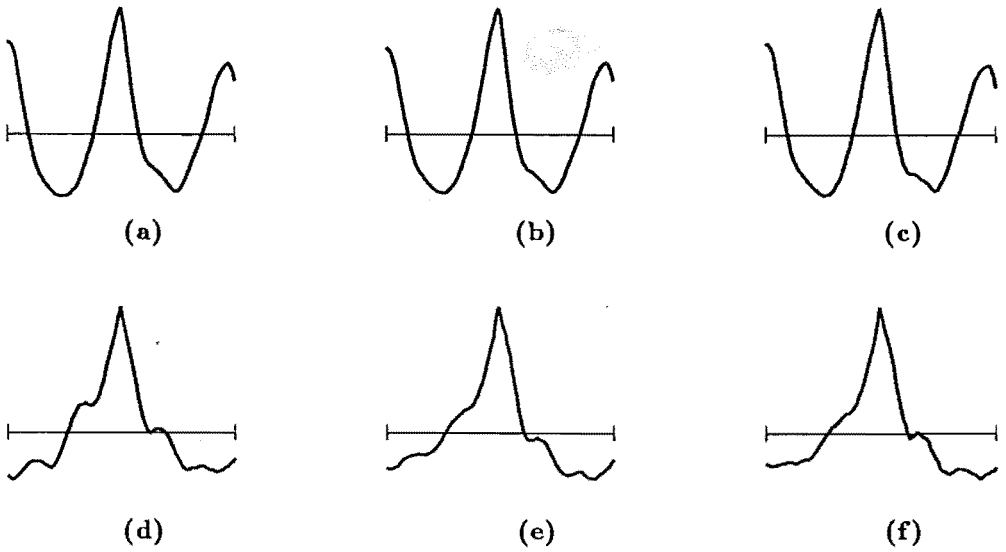


Figure 4.5. SAA signals obtained when the pitch estimate τ_p is varied, from 5.5ms to 30ms. In each case $\tau_s = 10$ms. Utterance TF-RAIN1: $\tau_p =$ **a:** 5.5ms, **b:** 10ms, and **c:** 30ms. Utterance WM-RAIN1: $\tau_p =$ **d:** 5.5ms, **e:** 10ms, and **f:** 30ms.

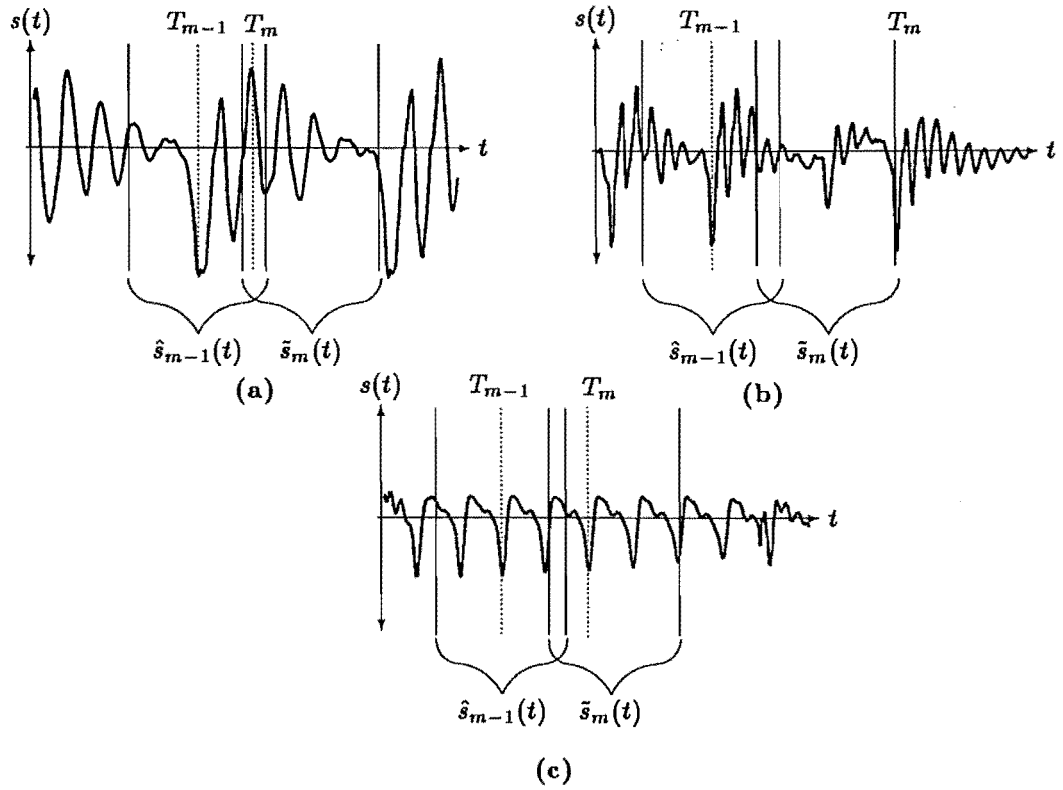


Figure 4.6. Segmentation errors produced by discrepancies between the pitch period p_m and the estimate of its average τ_p. **a:** m^{th} pitch period $p_m > \tau_p + \tau_s/2$, so that an erroneous peak is selected as the position of T_m. **b:** Signal maximum just outside extent of search segment, causing erroneous "peak" estimate. **c:** m^{th} pitch period $p_m < \tau_p - \tau_s/2$, causing the algorithm to skip a valid peak.

SAA process.

Since the results presented in this section indicate that the form of the SAA signal is insensitive to the values of τ_s and τ_p, I have set them equal to 12.8ms and 10ms respectively for all SAA processing described in the remainder of this thesis. Although these values are about twice as long as the average pitch period for utterances that are spoken by female speakers, the results presented in this section (notably in Fig.4.4*a*) indicate that in practice this has very little effect on the shape of the central portions of SAA signals from utterances spoken by female speakers. Note that a necessary constraint on the values of τ_s and τ_p is that $\tau_s < 2\tau_p$, so that $\tilde{s}_{m+1}(t)$ does not overlap $s(T_m)$.

4.2.4.2 Effects of differences between utterances

SAA extracts an estimate of the invariant component of a speech signal which, as defined for any particular utterance by (4.12), is an average over all segments in the utterance. Hence the exact form of the SAA signal for any particular utterance depends upon the content of that utterance. If the aim of performing SAA is to extract an estimate of a person's average glottal excitation, it is necessary that the content of the utterance be phonetically balanced (§4.2.2.1), so that the average vocal tract response can be considered negligible. How long the duration of an utterance should be to provide a reliable estimate of a person's glottal excitation depends on the content of the utterance. Computational experience suggests that 10 seconds is an adequate duration for a typical utterance. However, reliable SAA estimates can be obtained from utterances lasting no longer than 3 seconds, if they are carefully chosen to have appropriately balanced phonetic content.

In this section I present results which show the differences between SAA signals obtained from utterances of different durations and of different content. It is convenient here to introduce the quantity $\epsilon_{sa}^{(a:b)}$ defined by

$$\epsilon_{sa}^{(a:b)} = \int_{-\tau_s/2}^{\tau_s/2} \left[s_{sa}^{(a)}(t) - s_{sa}^{(b)}(t) \right]^2 \mathrm{d}t, \tag{4.22}$$

where $s_{sa}^{(a)}(t)$ and $s_{sa}^{(b)}(t)$ represent the SAA signals which are being compared. Both $s_{sa}^{(a)}(t)$ and $s_{sa}^{(b)}(t)$ are normalised to have a peak amplitude of unity. Throughout this thesis, $\epsilon_{sa}^{(a:b)}$ is invoked to quantify the differences between any pair of SAA signals. The labels (a) and (b) may therefore refer to actual utterances or to different forms of processing applied to a single utterance. Even though it merely identifies a difference between two estimates of the "true" invariant component $s^i(t)$, $\epsilon_{sa}^{(a:b)}$ is a useful measure of how "consistent" are the SAA signals obtained under different conditions.

Fig.4.7 shows the SAA signals obtained from sections of various durations taken from a single utterance. The number M of segments ranges from 10 for Fig.4.7*a* to 600 for Fig.4.7*f* (corresponding to utterance durations of about 100ms to 10s). The SAA signals corresponding to values of M greater than 200 are all seen to be very similar to each other. Curves of $\epsilon_{sa}^{(m:600)}$ versus m for utterances spoken by two different speakers appear in Fig.4.8. $\epsilon_{sa}^{(m:600)}$ is the error between the SAA signal for $M = 600$ and that for $M = m$. The two curves shown in Fig.4.8 confirm that the SAA signal does not change much for $M > 300$ for any particular utterance.

In order to illustrate the dependency of $s_{sa}(t)$ on the content of an utterance, Figs.4.9 shows SAA signals obtained from different utterances spoken by the same speaker. The utterances from which the SAA signals shown in Figs.4.9*a* and *b* were computed are of the same phrase, but spoken at different times, while the utterances

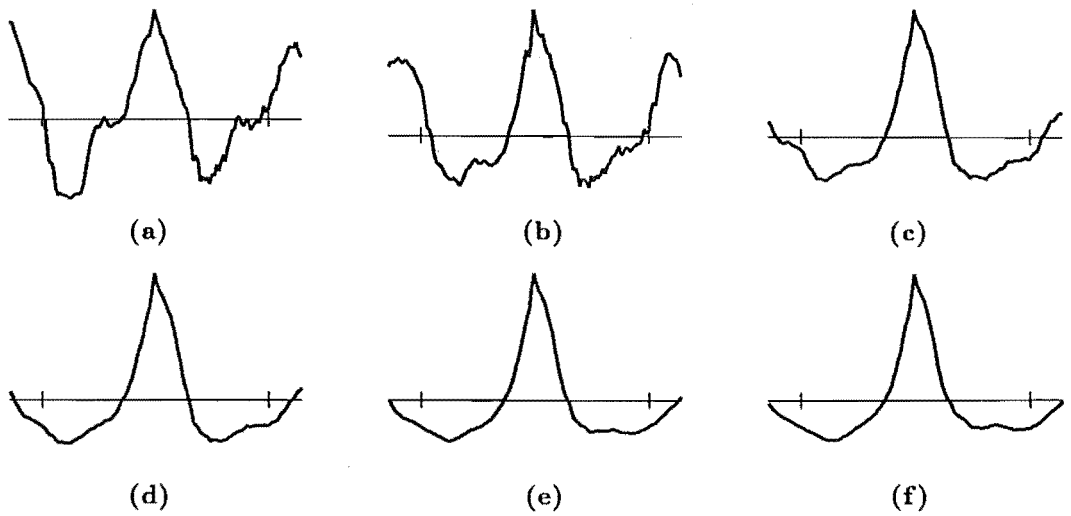


Figure 4.7. SAA signals obtained when the number of segments M is varied from 10 to 600. Utterance WM-BRIT (Table 1.2): **a:** $M = 10$, **b:** $M = 20$, **c:** $M = 100$, **d:** $M = 200$, **e:** $M = 400$, **f:** $M = 600$,

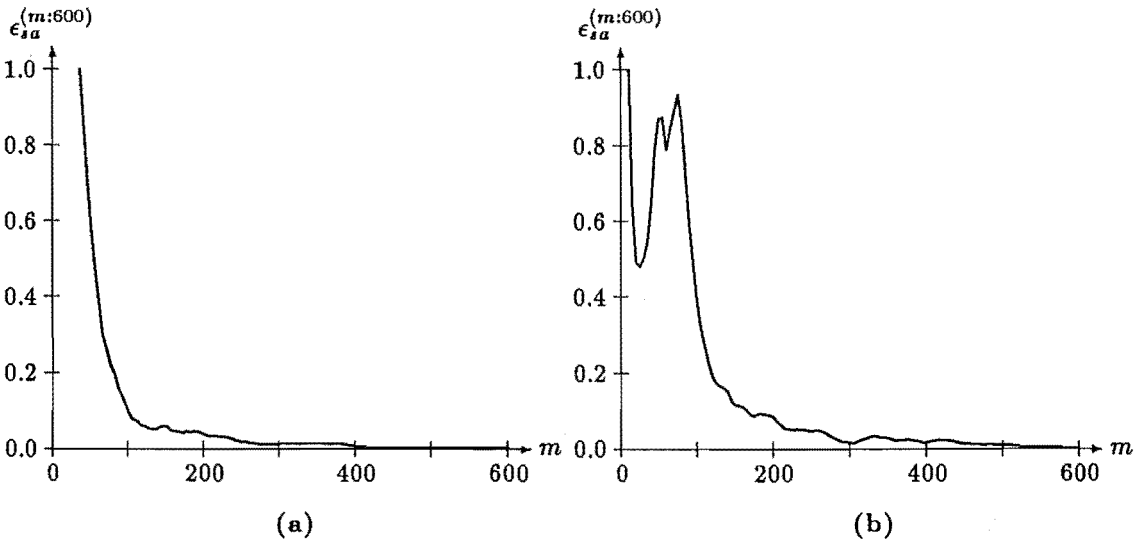


Figure 4.8. SAA error $\epsilon_{sa}^{(m:600)}$: **a:** Utterance WM-BRIT, and **b:** Utterance TF-WAL.

corresponding to Fig.4.9 *c–f* are of different phrases. Notice that there are more differences between the SAA signals for different phrases than there are between the SAA signals obtained from different occurrences of the same phrase. In addition, the SAA signals obtained from the "phoneme specific" phrases WM-VOWEL and WM-NASAL differ from the other, more phonetically "balanced", phrases (see §4.2.2.1).

Fig.4.10 shows SAA signals obtained from utterances spoken by different people. The results shown in Figs.4.9 and 4.10 suggest that, providing that the content of the utterance is relatively balanced, SAA signals obtained from utterances spoken by different people show greater diversity than SAA signals from utterances spoken by a single speaker. These results suggest that SAA could be useful as a technique of characterising a person's voice in a speaker recognition system (see §8.2.1.1).

The results shown in Figs.4.7 through 4.10 all pertain to utterances that are

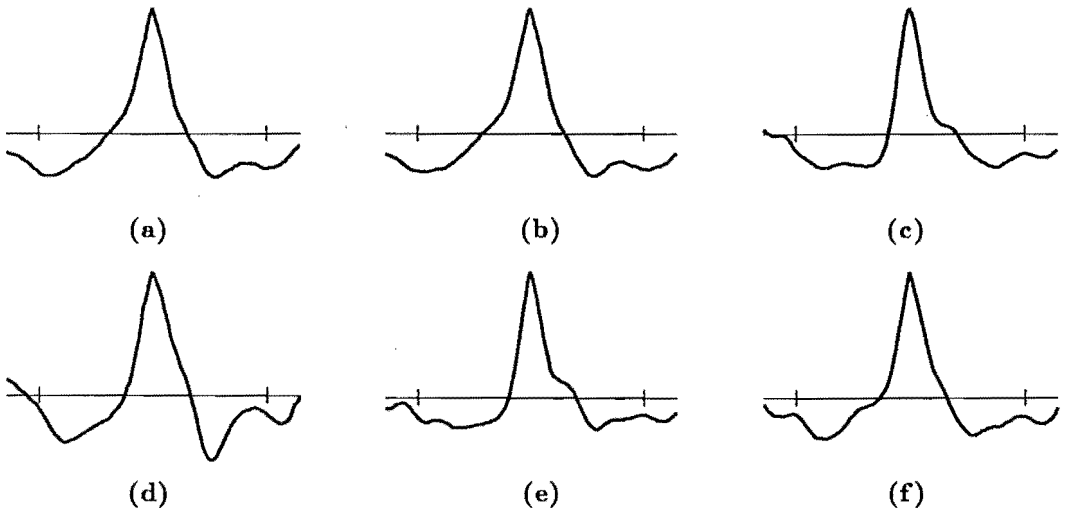


Figure 4.9. SAA signals from the voiced sections of different utterances spoken by a single person. Utterances (Table 1.2) **a:** WM-RAIN2, **b:** WM-RAIN3, **c:** WM-VOWEL, **d:** WM-NASAL, **e:** WM-TESTA **f:** WM-TESTB.

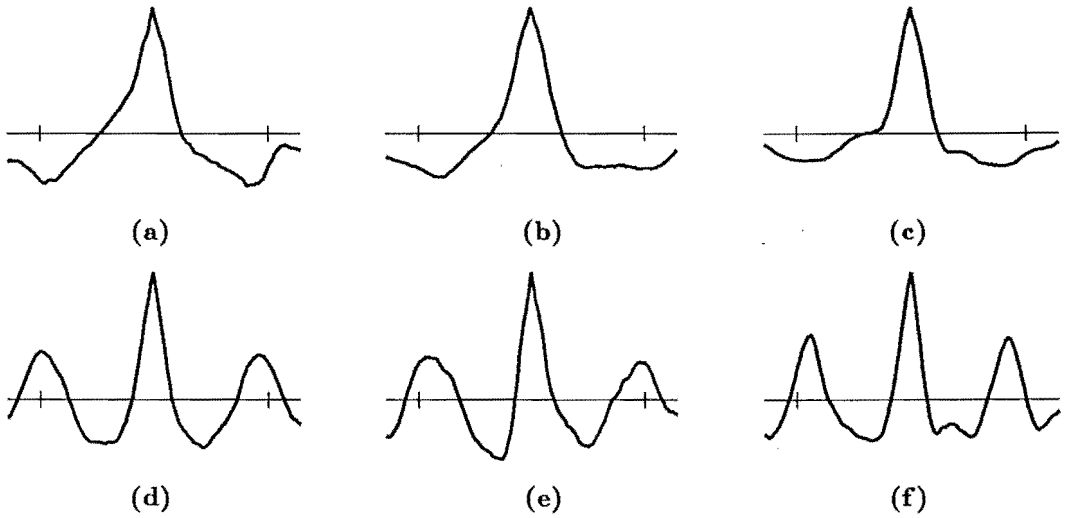


Figure 4.10. SAA signals from the voiced sections of an utterance spoken by different people. Utterances (Table 1.2) **a:** AM-WAL, **b:** WM-WAL, **c:** BM-WAL, **d:** TF-WAL, **e:** KF-WAL **f:** CF-WAL.

spoken in a "normal" voice. When a speaker talks in a different manner, the shape of the SAA signal changes markedly. Figs.4.11 *a*, *b*, and *c* respectively show the SAA signals obtained from utterances spoken with the muscles around the throat highly tensed, very relaxed, and in a normal speaking condition. Note the marked differences, especially between the SAA signal corresponding to the tense utterance and the other two. Part of this difference is due to the consequent change of pitch, but much of it also appears to be due to changes in the form of the glottal vibration caused by differences in the configuration of the muscles around the larynx.

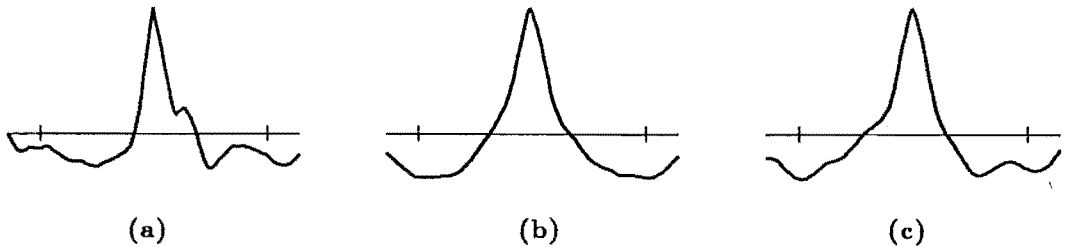


Figure 4.11. SAA signals obtained from utterances spoken in a: a "tense", b: a "relaxed", and c: a "normal" speaking manner. The utterances are (Table 1.2) a: WM-RAINT, b: WM-RAINR, c: WM-RAIN1.

4.2.4.3 Effects of phase distortion of the speech signal

Phase distortion of speech waveforms is of little consequence for most speech processing purposes because of the ear's insensitivity to such distortions (see §2.2.2.2). However, phase distortion changes the shape of the waveform, thereby affecting the locations of the peaks. In this section I show what effects phase distortion of a speech signal has on the shapes of the SAA signals obtained from that utterance. I then describe one method by which speech signals can be processed to partially correct for the ill-effects of phase distortion.

By phase distortion of a signal I mean that the magnitudes of the signal's spectral components are unchanged, while the phase of each component is altered. This occurs when the signal is filtered by a linear time-invariant system having a transfer function described by

$$H(f) = \exp(i\psi(f)) \tag{4.23}$$

with $\psi(f)$, which is called the phase distortion, being real and odd. Note that the case of $\psi(f) = \gamma f$, where γ is an arbitrary scaling constant, does not constitute phase distortion, since it merely represents a time delay through the system.

Graphic illustration of the effect of phase distortion on SAA is provided by the simple phase distortion $\psi(f) = \psi_o \text{ sgn}(f)$, where ψ_o is a real constant. Fig.4.12*a* shows a segment of speech, as it was recorded, while Figs.4.12*b* and *c* show the same segment after it has been subjected to phase distortions of $\psi(f) = \pi/6 \text{ sgn}(f)$ and $\psi(f) = \pi/2 \text{ sgn}(f)$ respectively. The segment of speech shown in Fig.4.12*d* has been distorted by a pseudo-random $\psi(f)$ which is composed of random numbers, with a uniform pdf, in the range from $-\pi$ to $+\pi$. The SAA signals shown in Figs.4.13*a,b,c* and *d* are obtained from the same utterances that the segments shown in Fig.4.12*a,b,c* and *d* are abstracted from.

As indicated by the differences between Figs.4.13*a,b,c* and *d*, phase distortion of the speech signal also changes the shape of the SAA signal. However, applying the inverse distortion to each SAA signal, as is shown in Fig.4.14, does not produce a signal like the $s_{sa}(t)$ of the undistorted speech signal. For minor amounts of phase distortion (such as is illustrated in Figs.4.12*a* and 4.13*b*), SAA produces a reasonably consistent estimate of $s^i(t)$. The moral of this story is that if one wishes to compare SAA signals obtained from different utterances (for instance for speaker recognition purposes, as described in §8.2.1.1), apparatus used for recording each utterance should have similar phase responses. This could be an important consideration for systems that accept speech over telephone lines, since the phase response may differ between each connection.

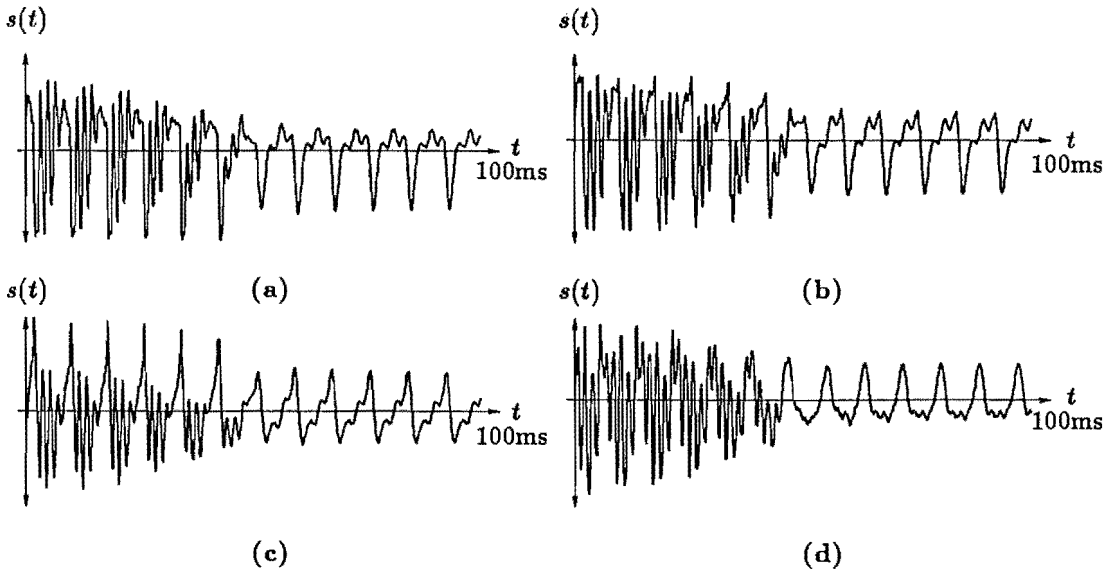


Figure 4.12. Phase distortion of a speech signal. **a:** Segment of signal as it was recorded. The same segment with a constant phase shift (4.23) of **b:** $\psi(f) = \pi/6 \text{ sgn}(f)$, **c:** $\psi(f) = \pi/2 \text{ sgn}(f)$, and **d:** $\psi(f) =$ pseudo-random noise.

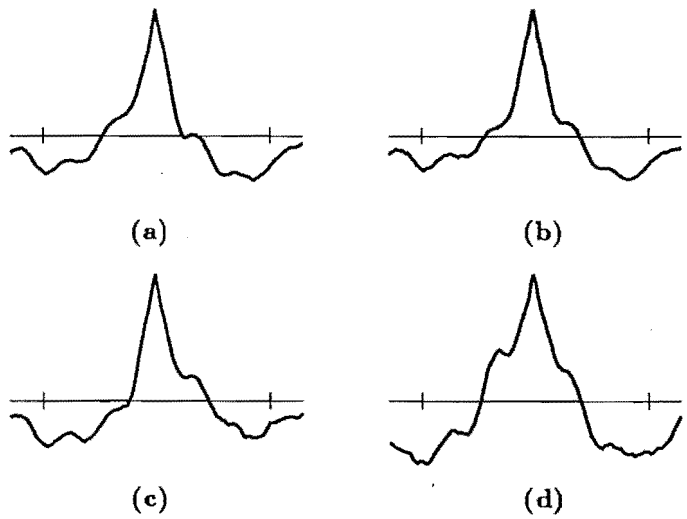


Figure 4.13. SAA signals of the utterance (see Table 1.2) AM-RAIN1 after it has been distorted by the phase distortions described in the text. **a:** No distortion, **b:** $\psi(f) = \pi/6 \text{ sgn}(f)$, **c:** $\psi(f) = \pi/2 \text{ sgn}(f)$, and **d:** $\psi(f) =$ pseudo-random noise.

One way to counteract the effects of phase distortion is to apply a "worse" distortion in the computer (!) which, when it is subsequently removed, removes the invariant phase distortion as well (cf. Bates, 1976; Bates and Robinson, 1982; Bates, 1982). By worse, I mean a distortion that changes the shape of the signal waveform much more than the distortion it is desired to repair, so that it dominates in the SAA processing (Dainty, 1973). In addition, a different distortion must be applied to each speech segment (or speckle image, §4.1), with the ensemble of such distortions having negligible mean. In the next paragraph I describe the technique which I employ to com-

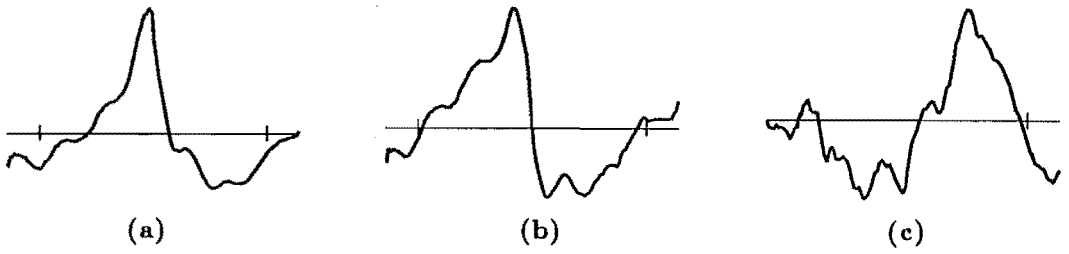


Figure 4.14. The SAA signals shown in Fig.4.13 after they have been processed by the inverses of the respective phase distortions that were applied to the speech signals. **a:** $\psi(f) = \pi/6\ \mathrm{sgn}(f)$, **b:** $\psi(f) = \pi/2\ \mathrm{sgn}(f)$, and **c:** $\psi(f)$ = pseudo-random noise.

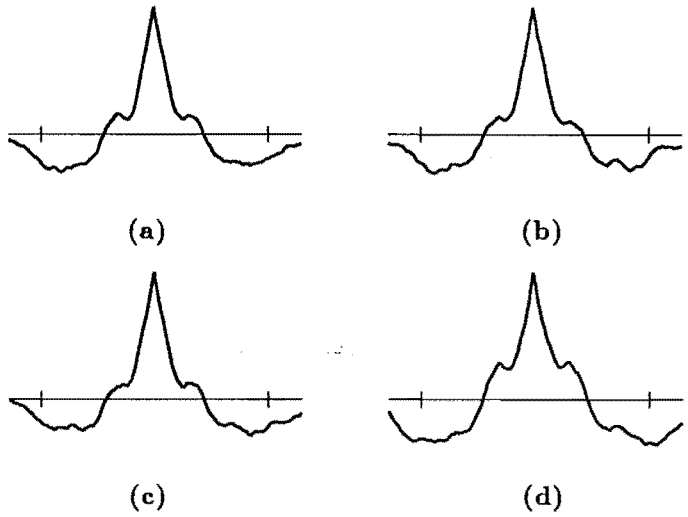


Figure 4.15. SAA signals of the utterances described in the caption of Fig.4.12, but where each speech segment is corrupted further with random noise before performing SAA. **a:** Undistorted, **b:** $\psi(f) = \pi/6\ \mathrm{sgn}(f)$, **c:** $\psi(f) = \pi/2\ \mathrm{sgn}(f)$, and **d:** $\psi(f)$ = pseudo-random noise.

pute the "noise-convolved" SAA signal $\zeta_{sa}(t)$ (where I use the term noise-convolved to mean the SAA signal obtained when each speech segment is distorted, in the computer, by random noise).

Onto each segment of speech $s_m(t)$, a segment of white noise $n_m(t)$, of equal duration to $s_m(t)$, is convolved. $n_m(t)$ is generated by a random number generator and has a uniform pdf. In this way a new ensemble of noise-convolved segments $\zeta_m(t)$ is produced:

$$\zeta_m(t) = s_m(t) \odot n_m(t), \quad -2\tau_-^a < t < 2\tau_+^a. \tag{4.24}$$

Shift-and-add is performed on the ensemble $\zeta_m(t)$ to form $\zeta_{sa}(t)$. Fig.4.15 shows the SAA signals of utterances, corresponding to these described in Fig.4.13, in which each segment has been convolved with noise before performing SAA. These are all much more similar than are the corresponding "ordinary" SAA signals shown in Fig.4.13. Fig.4.16 shows the mean square error $\epsilon_{sa}^{(u:d)}$ between the undistorted and each of the phase distorted SAA signals, computed for "ordinary" and "noise-convolved" SAA respectively. The much smaller values of $\epsilon_{sa}^{(u:d)}$ for the noise-convolved SAA signals

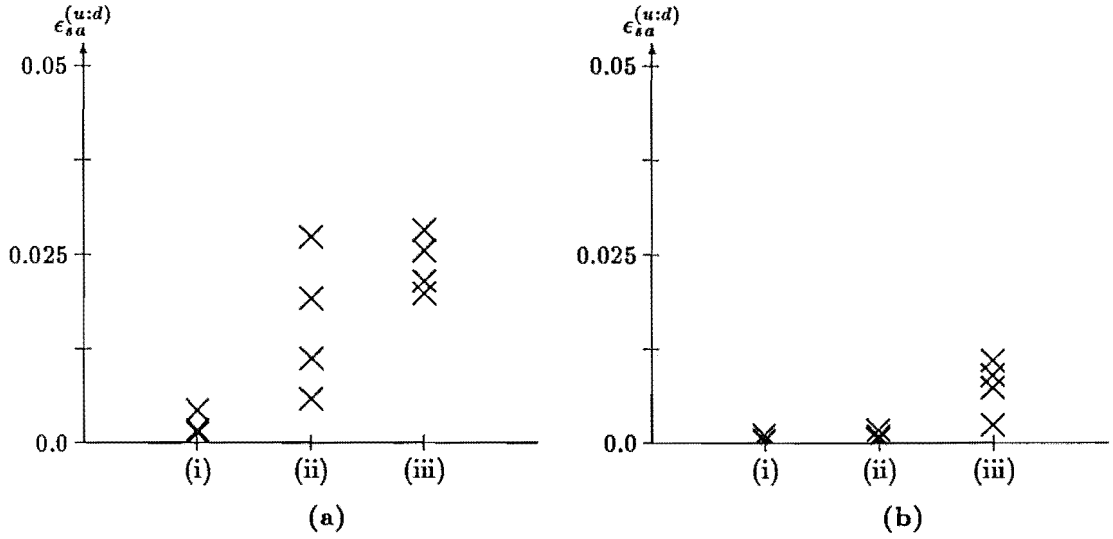


Figure 4.16. Mean square error $\epsilon_{sa}^{(u:d)}$ between undistorted and phase distorted SAA signals. **a:** "Ordinary" SAA and **b:** "noise-convolved" SAA. In each case the errors from the SAA signals of four different utterances (AM-RAIN1, AM-WAL, WM-RAIN1, TF-RAIN1) are shown. The labels on the abscissa of each plot identify the type of distortion suffered by the SAA signals represented in the column above. **i:** $\psi(f) = \pi/6$ sgn(f), **ii:** $\psi(f) = \pi/2$ sgn(f), and **iii:** $\psi(f)$ = pseudo-random phase distortion.

confirms that this noise distortion has "removed" at least some of the constant phase distortion that was applied to the signals.

Because of the increased computation required to produce the noise-convolved SAA signal, I have not investigated its use further than is described here. However, it is worthwhile to mention one implication of the results presented here. Since it appears that convolving white noise onto each segment before performing SAA tends to "remove" any constant phase distortion, one may ask what happens to the "phase" of $s^i(t)$ itself. Inspection of the SAA signals shown in Fig.4.15 shows that they are more symetrical than those appearing in Fig.4.13. This seems to confirm that the phase information in $s^i(t)$ has been partly "removed" by the noise-convolved SAA processing. This implies that the SAA signals obtained by this approach, while they are more consistent under different conditions of phase distortion, contain less information about $s^i(t)$ than do SAA signals obtained in the ordinary manner.

4.2.4.4 The effects of additive noise

Speech signals are often subjected to additive noise, which can take the form of transmission noise, interfering speech signals, or other environmental background sounds. In this section I examine how much the SAA signal of a given utterance is changed when the utterance is subjected to various levels of additive contamination. I investigate the effects of both (spectrally weighted) random noise and interfering speech signals.

As implied by the results of "noise-convolved" SAA presented in §4.2.4.3, SAA produces remarkably consistent results even when the speech signal is subjected to severe amounts of random (convolutional) distortion. This is emphasised by the SAA signals shown in Fig.4.17, which were obtained from speech signals corrupted by various levels of additive noise. The noise was generated by a random number generator and then filtered by a leaky first-order integrator having a feedback constant of 0.95. The noise therefore has a spectrum that falls off at about 6 dB per octave. At SNRs greater

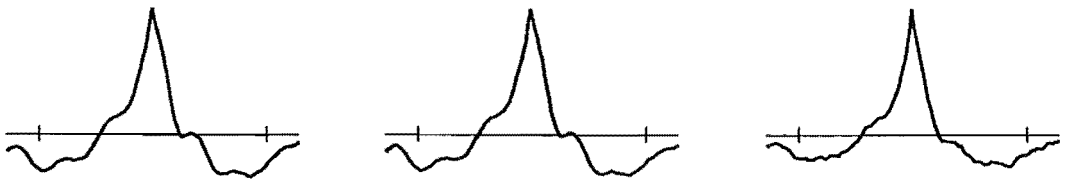


Figure 4.17. SAA signals of the utterance AM-RAIN1 that have been corrupted with various levels of additive noise. SNRs of **a:** 30 dB, **b:** 15 dB, **c:** 0 dB.

than 15 dB, the SAA signal is hardly different from the SAA signal obtained from the original utterance. At higher levels of additive noise, the SAA signal is more "noisy", and has a narrower "spike". Fig.4.18 shows the variation of $\epsilon_s^{(a:N)}a$, the error between the uncorrupted SAA signal and the SAA signal obtained from an utterance with SNR equal to NdB, for levels of SNR ranging from 30dB down to 0dB.

Contamination by interfering speech signals may be expected to have different effects on the SAA signal of an utterance than similar levels of contamination by "noise", since the interfering signal has itself an $s_{sa}(t)$ that is similar to the $s_{sa}(t)$ of the desired utterance. In order to investigate the effect of interference from other speech signals, I mixed two utterances in varying ratios, and performed SAA on the result. If the utterances are represented by $s^{(a)}(t)$ and $s^{(b)}(t)$, the mixed speech signal $s^{(\gamma)}(t)$ is given by

$$s^{(\gamma)}(t) = (1-\gamma)s^{(a)}(t) + \gamma s^{(b)}(t) \tag{4.25}$$

where γ is the mixing ratio, which is bounded by 0 and 1. The error $\epsilon_{sa}^{(a:\gamma)}$ is the error between the SAA signals of $s^{(a)}(t)$ and $s^{(\gamma)}(t)$. Fig.4.19*a* shows $\epsilon_{sa}^{(a:\gamma)}$ when the utterances labelled by (a) and (b) are spoken by different male speakers, while Fig.4.19*b* shows $\epsilon_{sa}^{(a:\gamma)}$ for the interference between a male and a female speaker. The errors are much greater in Fig.4.19*b* because the differences between the two SAA signals $s_{sa}^{(a)}(t)$ and $s_{sa}^{(b)}(t)$ are much greater than when the speakers are of different gender. These figures show that $s_{sa}^{(\gamma)}(t)$ follows a rather smooth, though not linear, transition between $s_{sa}^{(a)}(t)$ and $s_{sa}^{(b)}(t)$ as γ is varied between 0 and 1.

4.2.4.5 Dealing with unvoiced speech

In §4.2.1 through §4.2.4 I assume that unvoiced sections of each utterance are removed by standard VUV analysis techniques (see §3.1.2) before any of the SAA processing is performed. In this section I describe two methods by which the unvoiced sections of an utterance can be effectively removed, without performing an explicit VUV decision analysis. I also discuss the possibility of performing SAA on the unvoiced sections of an utterance.

A simple decision as to whether a segment of speech is voiced or unvoiced can be made by considering the peak signal amplitude within the segment. Most segments containing unvoiced speech have a peak magnitude considerably lower than that for a typical segment of voiced speech. Step 5 of the SAA algorithm can be modified so that any segment $s_m(t)$ for which

$$|s_m(T_m)| < \eta_{\text{uv}} \tag{4.26}$$

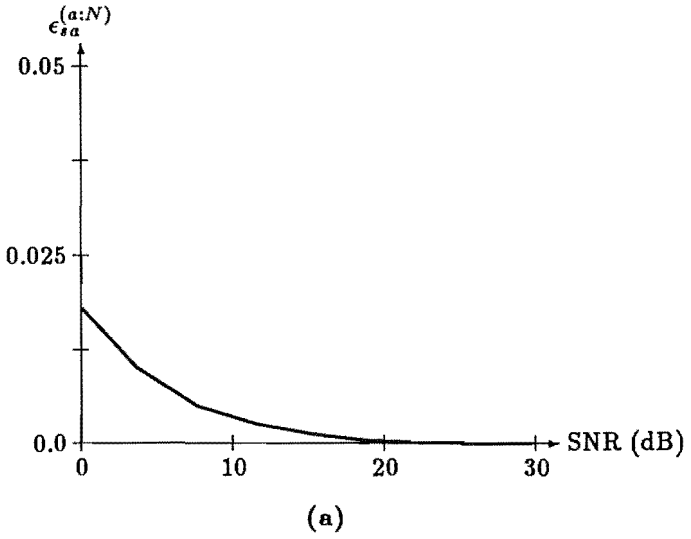


Figure 4.18. Error between SAA signals of uncontaminated utterance AM-RAIN1 and the same utterance subjected to various levels of additive noise.

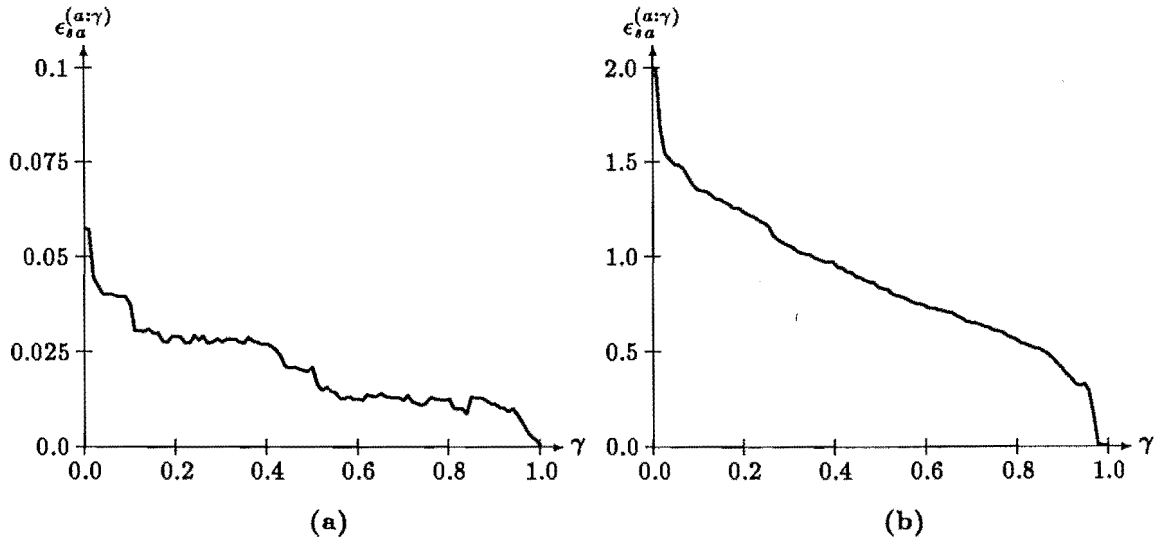


Figure 4.19. Error $\epsilon_{sa}^{(a:\gamma)}$ between SAA signals of uncontaminated utterances and signals composed of additive mixtures of the two utterances. **a:** Two male speakers (AM-RAIN1 and WM-RAIN1). **b:** Male and female speakers (AM-RAIN1 and TM-RAIN1). Note the different scales on the two graphs (necessary because of the much greater error for the male-female mixture).

is discarded, where η_{uv} is a threshold that is set to a value such that most of the unvoiced segments are excluded while most of the voiced segments are accepted (Elder, 19XX). Although application of (4.26) does not remove all the segments of unvoiced speech from an utterance (e.g. stop consonants usually exhibit high amplitude peaks), the averaging that is the essence of SAA implies that the few unvoiced segments which are not discarded do not have much effect on the final SAA signal.

Fig.4.20 shows SAA signals computed from the voiced sections only of the utterance AM-RAIN1, together with SAA signals computed from the entire utterance, with the threshold η_{uv} ranging from 25% down to 0% of the peak magnitude of the speech signal $s(t)$. The reduced value of M given in Fig.4.20 for each of the SAA signals

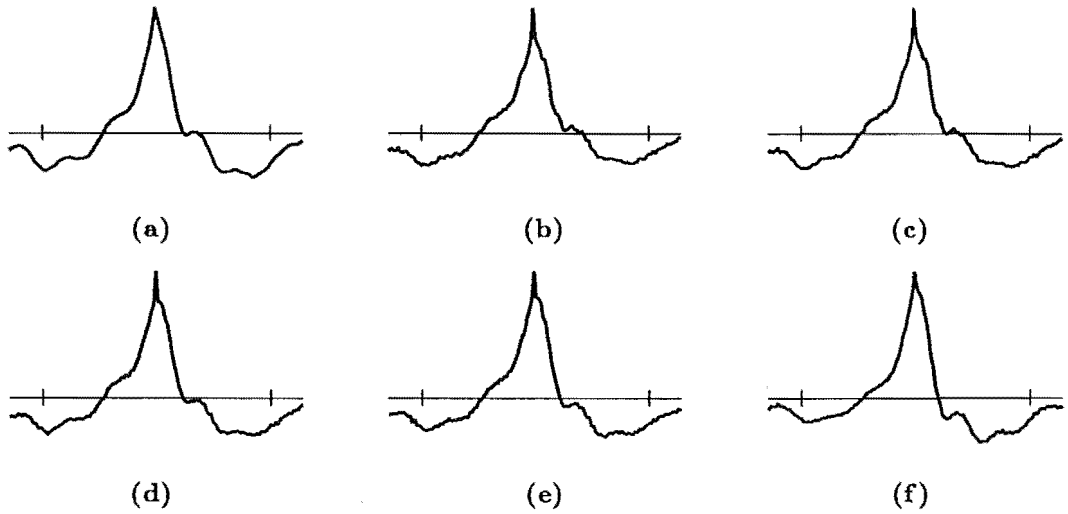


Figure 4.20. Demonstration of thresholding to exclude silent and unvoiced segments of speech signal. **a:** SAA of voiced segments of utterance AM-RAIN1 only (extracted by means of the VUV algorithm described in §3.1.2). SAA signals of the entire speech signal with the threshold η_{uv} set at **b:** 0, **c:** 1%, **d:** 5%, **e:** 10%, and **f:** 25% of the peak magnitude of $s(t)$.

when $\eta_{uv} > 0$ indicates how many of the segments in the utterance are excluded by the threshold. The main peak in the SAA signal appears to be "sharper" when η_{uv} is smaller, indicating that the main effect of unvoiced segments that are not excluded by the threshold is to increase the sharpness of the main peak. However, the experience of both my colleagues and myself (Elder, 19XX) suggests that the consistency, noted in §4.2.4.2, between SAA signals obtained from different utterances spoken by the same person, is also evident when the "threshold" method is invoked to discard the unvoiced speech segments. If the aim of SAA is to produce a "descriptor" of a speaker's average voice characteristics, as might be required for a speaker recognition system (§8.2.1.1), the threshold method appears to be adequate, while reducing the computational complexity of the processing that is required. However, if the aim is to characterise the average glottal excitation, the incorporation of a few unvoiced segments into $s_{sa}(t)$ may unacceptably distort the signal that is so produced. This question is examined further in §4.3.

Another method of effectively removing the unvoiced segments of speech is to low-pass filter the speech signal, since most of the energy in voiced sections of speech is concentrated in frequencies below about 3kHz (§2.1.4). This approach, which is discussed further in §5.4.3.3, is useful when the SAA processing is part of the speech encoding technique described in Chapter 5. This is because it separates the (effectively) voiced and unvoiced sections of speech without any of the difficulties of accurately performing a VUV decision analysis. It also means that sections of speech having a mixed excitation can be straightforwardly represented by a mixture of the two bands. Fig.4.22*a* shows the SAA signal obtained from the frequency band 0–2.5 kHz of the utterance AM-RAIN1. This is very similar to the SAA signal obtained from only the voiced sections of the unfiltered version of the same utterance (Fig.4.20*a*). Separating the speech signal into two sub-bands requires considerably more computation than either a straightforward VUV analysis or the simple unvoiced threshold described above, so I do not employ it for the results presented elsewhere in this chapter. However, it

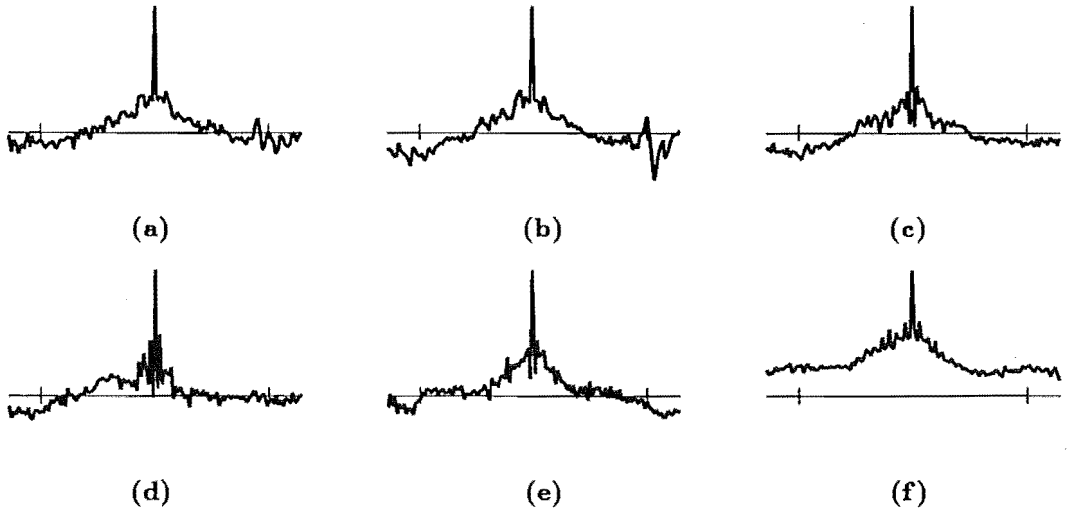


Figure 4.21. SAA signals computed for the unvoiced sections of the utterances **a:** AM-RAIN1, **b:** AM-RAIN2, **c:** AM-WAL, **d:** WM-RAIN1, **e:** TF-RAIN1, and **f:** TF-WAL.

reappears in Chapter 5, where the extra computation is justified on the grounds of the consequent improvement in performance of the CLEAN method of speech analysis.

The reasoning underlying the development of SAA processing of voiced speech can also be applied to the unvoiced sections of an utterance. In §4.2.1, each segment of (voiced) speech is described by a convolution between an invariant component (that is dominated by the glottal pulse shape) and a variant component (that mainly represents the vocal tract filter). The SAA signal is thought of as approximating the invariant component. Unvoiced speech can be described by a similar convolution, with the invariant component representing the "average" unvoiced excitation. SAA of the unvoiced sections of an utterance should therefore produce an estimate of this average unvoiced excitation.

Fig.4.21 shows SAA signals obtained from the unvoiced sections of several utterances. These SAA signals are very different from the SAA signals computed from the voiced sections of the same utterances (Fig.4.9). The narrowness of their main peaks can indicate either that the unvoiced excitation contains a broad-band invariant component, or that it varies so much between segments that its invariant component (as defined in §4.2.1) is negligible. Note that the second alternative does not imply the first, since an ensemble of narrow-band signals $\nu_m(t)$ can have an impulsive SAA signal if the ensemble $\langle |N_m(f)| \rangle_m$, where $N_m(f) = \mathcal{F}\{\nu_m(t)\}$, is flat (see §4.3.2.2). In §5.4 I present further results indicating that the SAA signal obtained from the unvoiced sections of an utterance can be thought of as representing the "invariant component" of the unvoiced excitation for that utterance.

Note the similarity between the SAA signal of the high frequency sub-band 2.5–5 kHz shown in Fig.4.22*b* and the SAA signal of the unvoiced sections of the same utterance shown in Fig.4.21*a*. This again illustrates the point made above about the usefulness of separating the speech into sub-bands as a means of separately characterising the essentially voiced and unvoiced parts respectively of a speech signal.

Figure 4.22. SAA signals obtained from sub-bands of the utterance AM-RAIN1. **a:** SAA signals of the sub-band 0–2.5kHz. **b:** SAA signal of the sub-band 2.5–5kHz. **c:** SAA signal of the sub-band 2.5–5kHz after it has been frequency shifted so that it lies in the band 0–2.5kHz.

4.2.4.6 To normalise or not to normalise

The loudness of speech varies a great deal between different syllables and words (§3.1.1). Hence the question arises as to whether or not the segments $s_m(t)$ should be normalised by dividing each one by $s_m(T_m)$ before forming $s_{sa}(t)$. Such normalisation causes all segments to be weighted equally, in the sense that their relative amplitudes are discarded. However, it also means that segments of low amplitude, which one would expect to have a higher level of "noise" corruption, are amplified relative to the segments with high signal amplitudes.

Fig.4.23 shows SAA signals computed from the voiced sections of an utterance with (Fig.4.23*a*) and without (Fig.4.23*b*) amplitude normalisation. Figs.4.23*c* and *d* show SAA signals, with and without normalisation respectively, computed from the entire utterance, with a threshold $\eta_{uv} = 10\%$ of the peak amplitude of $s(t)$. The "noise" exhibited by the SAA signal shown in Fig.4.23*c* is the result of the amplification of the unvoiced segments of the utterance that are not excluded by the threshold.

Since the amplitude variations between segments are modelled by the variant component $s_m^v(t)$ of the speech model (§4.2.2.1), it seems reasonable that they should not contribute to the SAA signal, which is a model of the invariant component $s^i(t)$ (§4.2.1). For this reason I prefer to normalise each segment $s_m(t)$ before averaging them to form $s_{sa}(t)$. I rely on the threshold η_s or η_{uv} to exclude segments of low amplitude which may contain inter-word silences or unvoiced speech.

4.2.4.7 Differentiating the speech signal to improve SAA

As described in §3.2.1, the spectrum of the glottal waveform falls off at approximately 12dB/octave, while the effect of lip radiation is to apply an approximate +6dB/octave emphasis to the speech spectrum. The "effective" glottal excitation is therefore approximately modelled with a -6dB/octave spectral slope. By pre-emphasising the speech signal with a first-order differentiation, the spectrum of the effective glottal excitation becomes approximately flat. This means that each glottal pulse is nearly "impulsive", which, as mentioned in §4.2.2.2, improves the performance of the SAA algorithm by reducing the ghosting due to wrongly-identified peaks.

Fig.4.24 shows SAA signals computed from the voiced sections of utterances that have been pre-emphasised by first order differentiation. Integrating the SAA signals shown in Fig.4.24 results in the signals shown in Fig.4.25. The differences between these signals and the SAA signals of the same, but undifferentiated, utterances (Fig.4.9), emphasise the non-linearity of SAA (because of the signal dependence of ghosting, see §4.1.2). However, comparing the signals in Figs.4.25 and 4.9 with the

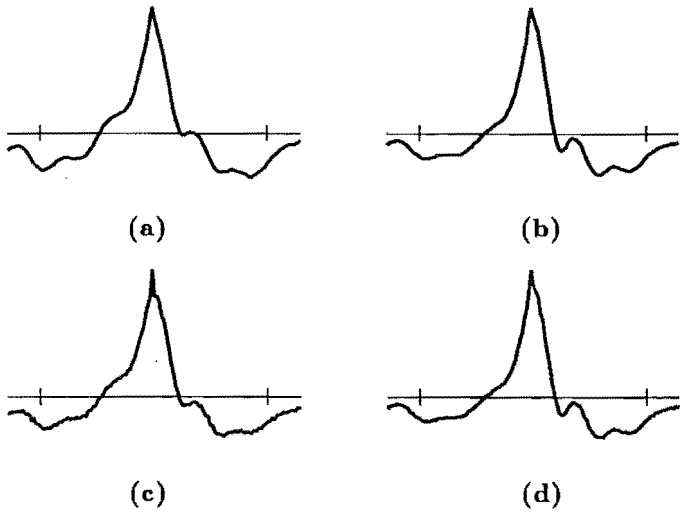


Figure 4.23. Differences between SAA signals when the segments are normalised or un-normalised. SAA signals computed from **a:** normalised, and **b:** un-normalised, segments of the voiced sections only of utterance AM-RAIN1. SAA signals of the **c:** normalised, and **d:** un-normalised, segments of the entire utterance, with a threshold set to 10% of the peak amplitude of $s(t)$ to exclude silent and unvoiced segments.

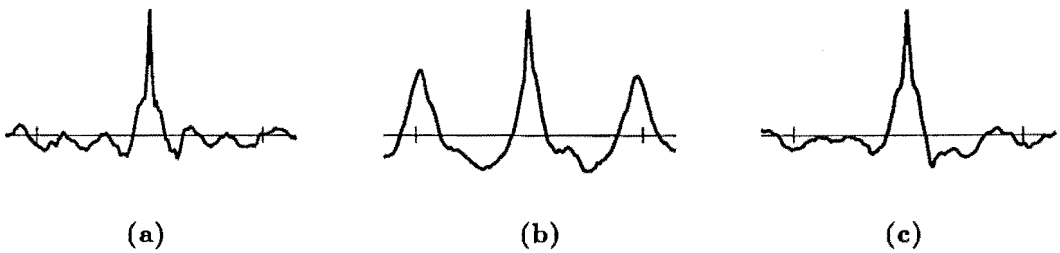


Figure 4.24. SAA signals from the voiced sections of utterances that have been pre-emphasised with a first-order differentiation. Utterances **a:** AM-RAIN1, **b:** TF-RAIN1, and **c:** WM-RAIN1.

waveforms obtained by other glottal estimation techniques (§3.4.1) indicates that SAA may give an improved estimate of the glottal pulse if the speech signal is first pre-emphasised. This is discussed in more detail in §4.3.

Note that if the speech signal is pre-emphasised before performing SAA, the threshold method described in §4.2.4.5 of excluding the unvoiced sections of speech is not so effective. This is because pre-emphasis enhances the speech amplitude during the unvoiced sections relative to the amplitude during voiced sections. For this reason it seems to be best to perform pre-emphasis only when the speech signal has been separated into voiced and unvoiced sections, or into high and low frequency sub-bands, as described in §4.2.4.5.

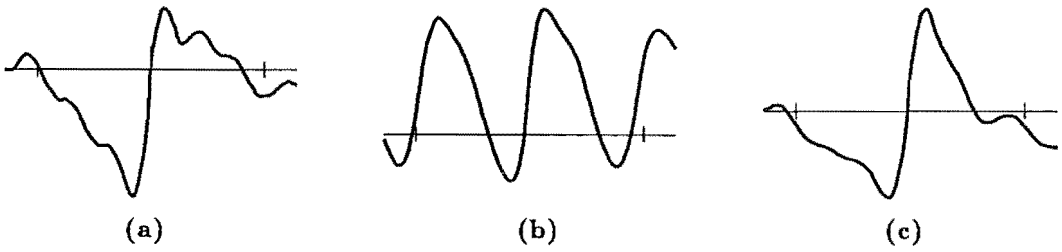


Figure 4.25. Signals resulting from first-order integration of the SAA signals shown in Fig.4.24. Each part corresponds to the same part of Fig.4.24.

4.3 Relating the SAA signal to the glottal waveform

As §4.2 explains, SAA processing of speech signals produces an estimate of the invariant component (i.e. the component that is the same in each pitch period, see §4.2.1) of a speech utterance. However, such an estimate contains contributions from both the glottal excitation and vocal tract components that together make up the traditional source-filter model of speech (cf. §2.3.1.4). In this section I investigate the relationship between the SAA signal and the glottal excitation component of the source filter model. In particular, I investigate the validity of the assumption that $s_{sa}(t)$ is an estimate of the invariant component $g^i(t)$ of the glottal excitation waveform. §4.3.1 compares the results obtained by SAA with those obtained by other methods of estimating the glottal pulse. In order to examine the extent to which SAA ghosting distorts the estimate of the glottal excitation, §4.3.2 presents the results of performing SAA processing on synthetic speech, where the form of the glottal excitation is known. In §4.3.3 I introduce a method which attempts to refine the SAA signal to obtain an improved estimate of the "true" glottal pulse. Finally, §4.3.4 presents some conclusions about what the relationship between the SAA signal and the glottal waveform really is.

4.3.1 Results from real speech

As implied in §4.2.2.2, $s_{sa}(t)$ is actually the first derivative of the glottal flow "pulse", because of the differentiation effect of lip radiation. Furthermore, if the speech is pre-emphasised before performing SAA, in the manner described in §4.2.4.7, $s_{sa}(t)$ is approximately the second derivative of the glottal flow. Unfortunately, integrating the SAA signal to counteract these differentiations can result in low frequency instability because integration introduces a pole at zero frequency. In order to overcome the low frequency instability, a "leaky" integrator can be employed. in the results presented here, I only invoke a leaky integrator when performing the double integration on the SAA signals obtained from pre-emphasised speech (§4.2.4.7). I use a double integrator of the form

$$H(z) = \frac{1}{(1 - z^{-1})(\alpha - z^{-1})} \tag{4.27}$$

where $\alpha = 0.95$ is the leak constant.

In addition to the instability caused by integration, the zero that is introduced at d.c. by the differentiation of lip radiation means that the estimated glottal flow waveform cannot be referenced to a true zero value. This is a difficulty inherent in almost all methods of estimating the glottal flow from the speech signal (§3.4.1).

In §4.3.1.1 I present results which compare estimates of the glottal excitation obtained by SAA processing with those obtained by the inverse filtering technique of

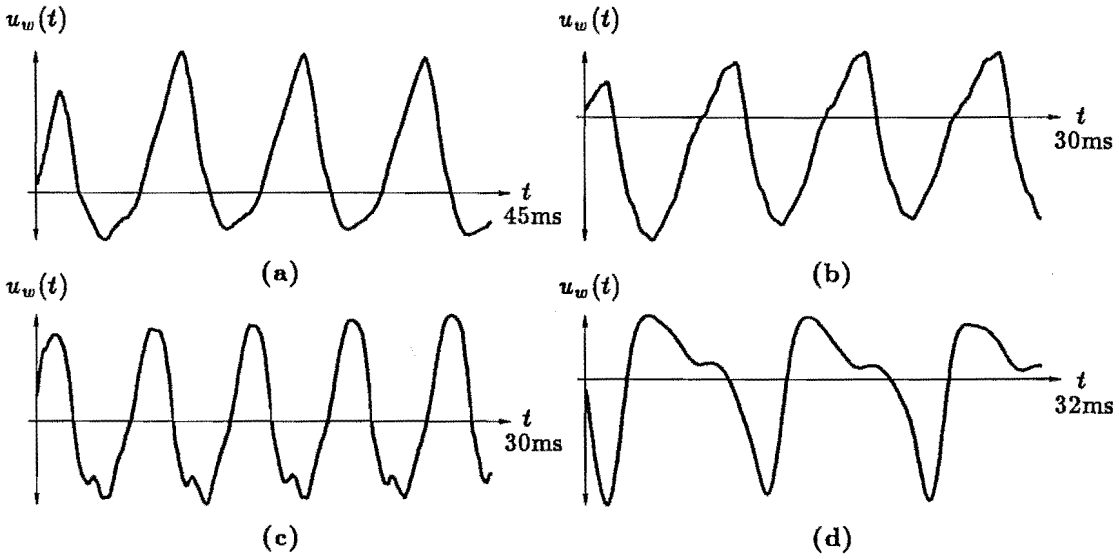


Figure 4.26. Typical inverse filtering waveforms obtained from four segments of the utterance AM-RAIN1 by the method of Wong *et al.* (1979). See §3.4.1 for details on the inverse filtering technique.

Wong *et al.* (1979), while in §4.3.1.2 I compare them with the results of long term spectral averaging of the voiced speech signal (cf. Boves, 1984).

4.3.1.1 Comparison with inverse filtering

SAA produces an estimate of the average shape of the glottal excitation over an entire utterance, whereas the inverse filtering techniques described in §3.4.1 produce an estimate of the glottal waveform on a cycle-to-cycle basis. Fig.4.26 shows glottal waveforms obtained from several different sections of the utterance AM-RAIN1 by the inverse filtering method of Wong *et al.* (1979). These glottal waveform estimates are different, suggesting that the shape of the glottal waveform changes as different sounds are uttered. Fig.4.27*d* shows the *average* glottal waveform of the utterance AM-RAIN1. This is obtained by dividing the voiced sections of the utterance into sections of 50ms in duration, performing the inverse filtering operation on each section (thereby obtaining an ensemble of signals such as those shown in Fig.4.26), and synchronously (according to the SAA paradigm) averaging the ensemble so formed. The similarities between this signal and the (integrated) SAA signal shown in Fig.4.27*b* support the contention that SAA processing produces an estimate of the average glottal excitation.

In §4.2.4.7 I mentioned that differentiating the speech signal before performing SAA should reduce ghosting in the SAA signal because it flattens the spectrum of the glottal pulse. Fig.4.28*a* shows the SAA signal computed from the (voiced sections of the) differentiated utterance AM-RAIN1. The twice-integrated "equivalent glottal-flow" signal corresponding to this appears in Fig.4.28*c*. Like the "undifferentiated" SAA signal, this also is quite similar to the average glottal excitation shown in Fig.4.27*d*.

Because of the overall similarity between the SAA signals and the inverse filter estimates of the glottal waveform, as illustrated by the results shown in Figs.4.26 to 4.28, I am encouraged in my assumption that the SAA signal represents the average glottal excitation (§4.2.2). The differences that are apparent between the signals shown in Figs.4.27*a,b*, and 4.28*b* may be due to the different assumptions made in their derivations, which lead to different types of errors in the estimates. For instance, the

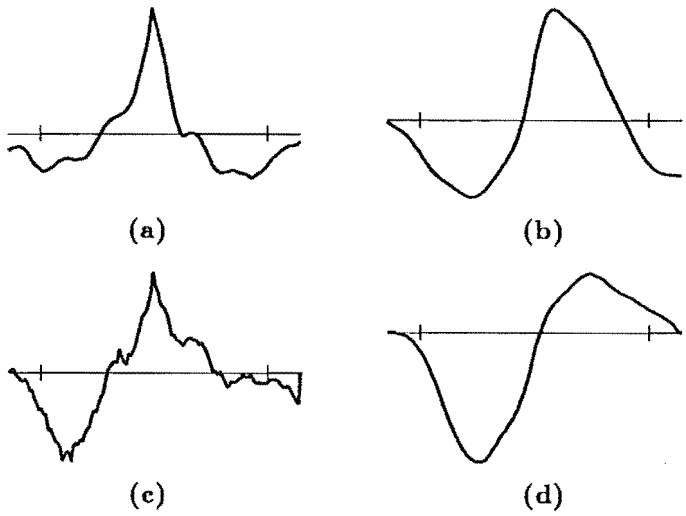


Figure 4.27. Comparison between SAA and inverse filtering as techniques of characterising the glottal excitation. **a:** SAA signal of the voiced sections of the utterance AM-RAIN1. **b:** Integrated SAA signal. **c:** Average glottal excitation signal obtained by performing SAA on the ensemble of signals produced by inverse filtering contiguous 100ms segments of the voiced sections of the utterance AM-RAIN1 (Note that SAA is performed on signals which are actually differentiated versions of the examples shown in Fig.4.26). **d:** Integrated version of the signal shown in *b*.

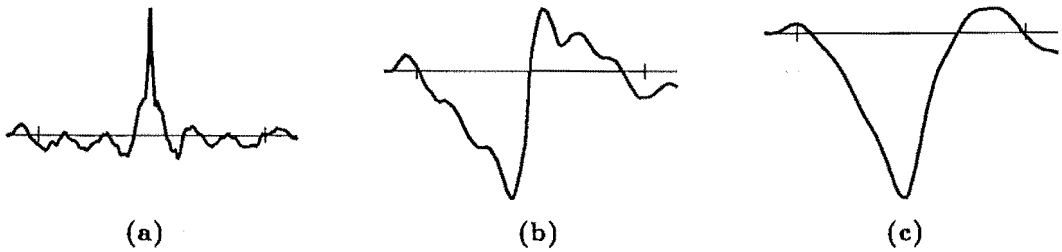


Figure 4.28. **a:** SAA performed on the utterance AM-RAIN1 after it has been pre-emphasised by first-order differentiation. **b:** Integrated, and **c:** twice-integrated, versions of the SAA signal shown in *a*.

inverse filtering approach which produced the signal shown in Fig.4.27*d* assumes that the vocal tract is modelled by an all-pole filter, whereas the SAA technique which gave rise to the signals shown in Figs.4.27*a* and 4.28*b* assumes that, if the SAA signal is to represent the average glottal waveform, the long term average vocal tract filter response is negligible. Since actual speech signals violate these assumptions in different ways, one expects the errors in the two glottal excitation estimates to also differ.

4.3.1.2 Comparison with the Long-term average spectrum (LTAS)

Since SAA produces an estimate of the average glottal pulse shape, it is relevant to compare SAA signals with the results obtained by the long-term average spectrum (LTAS) approach. The LTAS of an utterance is obtained by dividing it into many short segments, computing the power spectrum of each, and then obtaining the ensemble

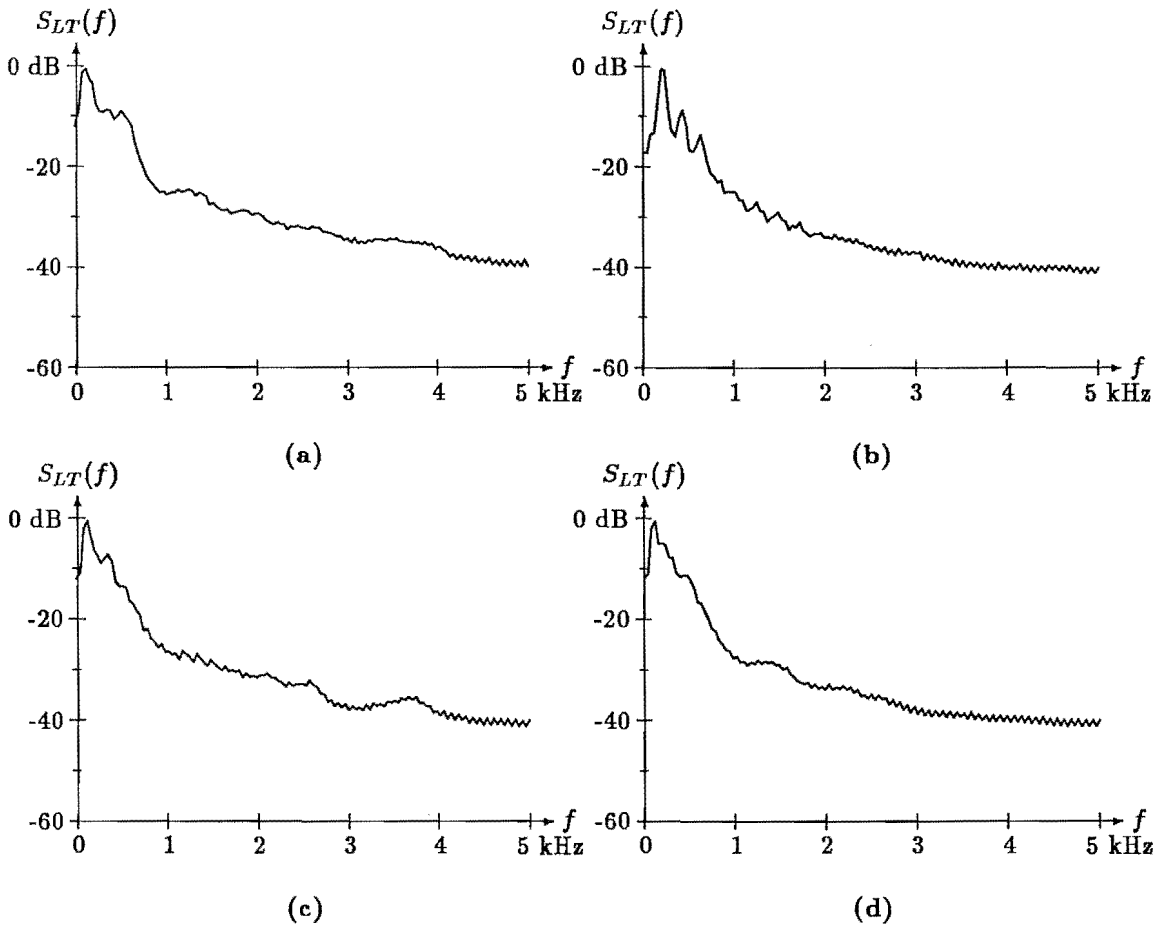


Figure 4.29. LTAS obtained from the voiced sections of several utterances. A segment length of 12.8ms, and a Blackman window, were employed in their computation. Utterances **a:** AM-RAIN1, **b:** TF-RAIN1, **c:** AM-WAL, and **d:** WM-WAL.

average. It is thus the spectral equivalent of SAA. Stockham *et al.* (1975) employ spectral averaging to estimate the invariant component of the distortion introduced by acoustic recording equipment on old sound recordings (§1.3.2). Boves (1984) compares LTAS with the average spectra of actual glottal waveforms, concluding that the LTAS of an utterance has similar characteristics to the average glottal excitation spectrum, with at least some of the observed differences arising because first-order differentiation is an inadequate model of lip radiation (Boves, 1984, §5.3.5).

Fig.4.29 shows the LTAS obtained from voiced sections of several different utterances. The spectra of the corresponding SAA signals are shown in Fig.4.30. These figures indicate that the SAA spectra, although they are of similar overall shape, fall off much faster at high frequencies than do the LTAS. The greater smoothness of the LTAS occurs because of the spectral averaging (cf. §1.3.1.2) over segments with different pitch frequencies. This means that the pitch harmonics tend to smooth out. By contrast, the SAA spectra are obtained from a single instance of a quasi-periodic (albeit of approximately only one period in duration) signal, so that they contain prominent harmonic structure.

The much sharper roll-off at high frequencies in the SAA spectra when compared to the LTAS probably results from the low frequency components of the speech signal dominating in the SAA processing. The high frequency components, which are not in general synchronised with the low frequency components (cf. Fujimura, 1968) tend to cancel out in the SAA averaging (§5.4.3). The spectral averaging, however, is

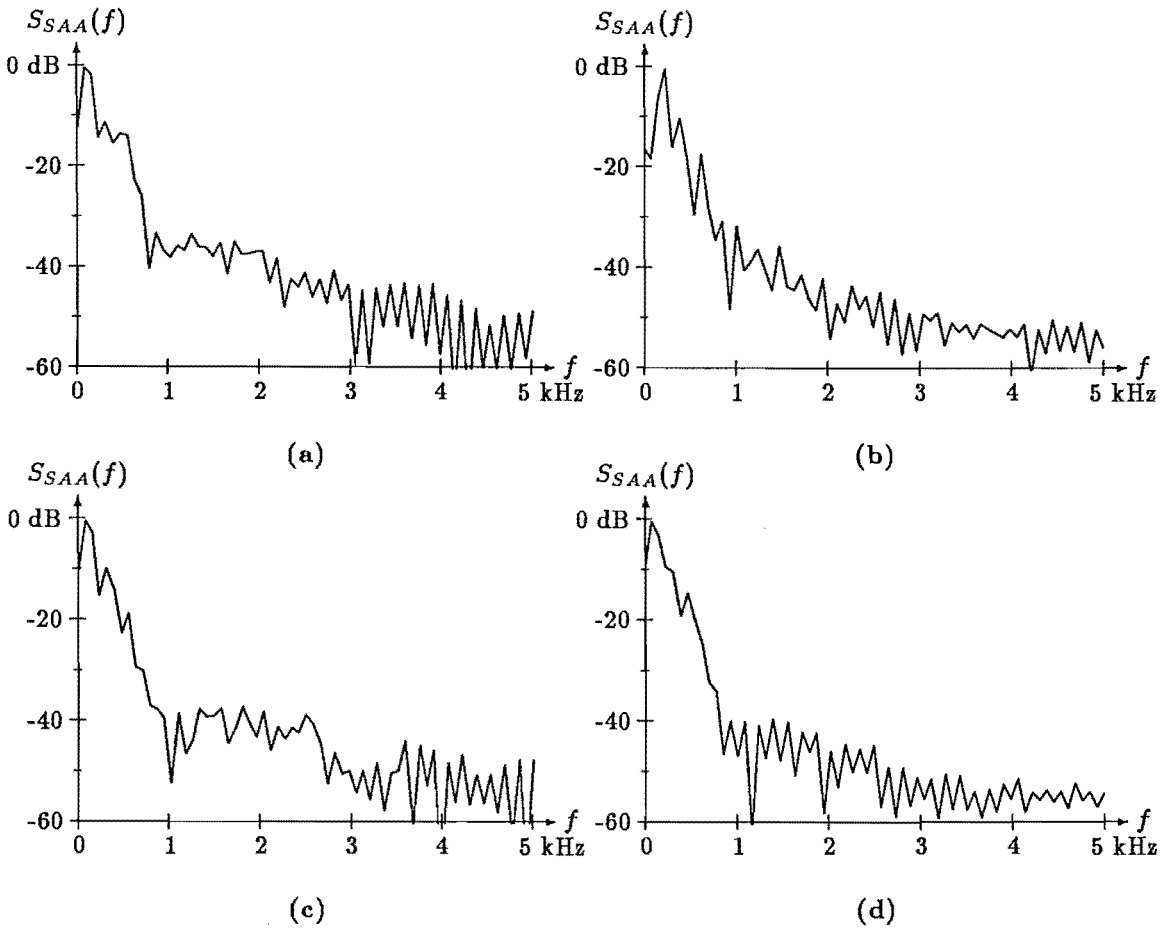


Figure 4.30. Spectra of SAA signals obtained from several utterances. The utterances pertaining to the spectra shown in a:, b:, c:, and d: are the same as those described in Fig.4.29.

non-coherent, so all such components accumulate in the sum.

4.3.2 Results from synthetic speech

In order to determine the efficacy of SAA at extracting the true glottal pulse from a speech utterance, SAA processing was performed on synthetic speech generated with a fixed excitation pulse shape. In §4.3.2.1 LPC parameters from actual utterances are employed as the filter component of the synthetic speech, while in §4.3.2.2 an artificial filter, having negligible invariant component, is employed.

4.3.2.1 Synthetic speech generated from actual LPC parameters

One way of generating synthetic speech with a known excitation is to excite the LPC filter coefficients obtained from actual utterances with a fixed pulse shape. It is preferable to employ "actual", rather than random, LPC filter coefficients because the signals generated from random coefficients are not really "speech-like" and so behave differently in the SAA processing. Fig.4.31 shows the synthetic pulse shapes used as the excitation signals in the various trials presented here. Speech was generated by convolving the impulse response of an LPC filter with an excitation waveform made up of copies of the synthetic pulse, positioned according to the pitch information from the same utterance from which the LPC filter parameters were obtained. Table 4.2 lists the synthetic utterances, detailing the pulse shapes and the natural utterances from which the speech parameters were obtained. Note that only the voiced sections of speech were

(a) (b)

Figure 4.31. Synthetic glottal pulse shapes. **a:** Triangle "saw-tooth". **b:** Raised cosine or sine-squared.

Label	Pulse shape	Actual utterance
AM-SAW	saw-tooth	AM-RAIN1
AM-SIN	sine-squared	AM-RAIN1
AM-IMP	impulse	AM-RAIN1
WM-SAW	saw-tooth	WM-RAIN1
WM-SIN	sine-squared	WM-RAIN1
WM-IMP	impulse	WM-RAIN1
TF-SAW	saw-tooth	TF-RAIN1
TF-SIN	sine-squared	TF-RAIN1
TF-IMP	impulse	TF-RAIN1

Table 4.2. List of the synthetic utterances that were generated to test the efficacy of SAA at extracting the actual glottal pulse shape. The "Pulse shape" column refers to the shape of the excitation employed in the synthesis while the "Actual utterance" column refers to the utterance from which the LPC parameters were abstracted.

retained for this experiment. The autocorrelation method was invoked to compute 10 LPC coefficients every 10ms during the utterance. The speech samples in each analysis frame, which were of 20ms in duration, were multiplied by a Hamming window before LPC analysis was performed (see §3.2 for further details on LPC analysis techniques).

SAA was performed on each of the synthetic utterances listed in Table 4.2, resulting in the SAA signals shown in Fig.4.32*a* through Fig.4.32*f*. The shapes of the SAA signals are by and large similar to the shapes of the original pulse shapes shown in Fig.4.31. However, significant distortions are apparent for some of the utterances, especially in the interval immediately after the peak of the saw-tooth pulse (Figs.4.32*a*,*b*). This distortion could arise either because of ghosting in the SAA process, or because the LPC filter has an invariant component — in the sense that its ensemble average is not negligible (see §4.2.1). Fig.4.33 shows the SAA signals obtained from impulse-excited synthetic speech signals. These SAA signals provide estimates of the invariant component $v^i(t)$ of the vocal tract filter, as far as it is modelled by the LPC coefficients. Note that deconvolving these estimates of $v^i(t)$ from the $s_{sa}(t)$ shown in Fig.4.32 does not restore the shapes of the original pulses (see Fig.4.34). This is because the ghosting is dependent on the actual form of the signal, so is different for differently excited synthetic utterances. In §4.3.2.2 I describe the use of synthetic utterances, that are generated in such a way that their invariant component is negligible, to investigate the amount of ghosting that can be expected when performing SAA on speech-like signals. §4.3.3 describes a method by which the distortions, whether due to SAA ghosting or

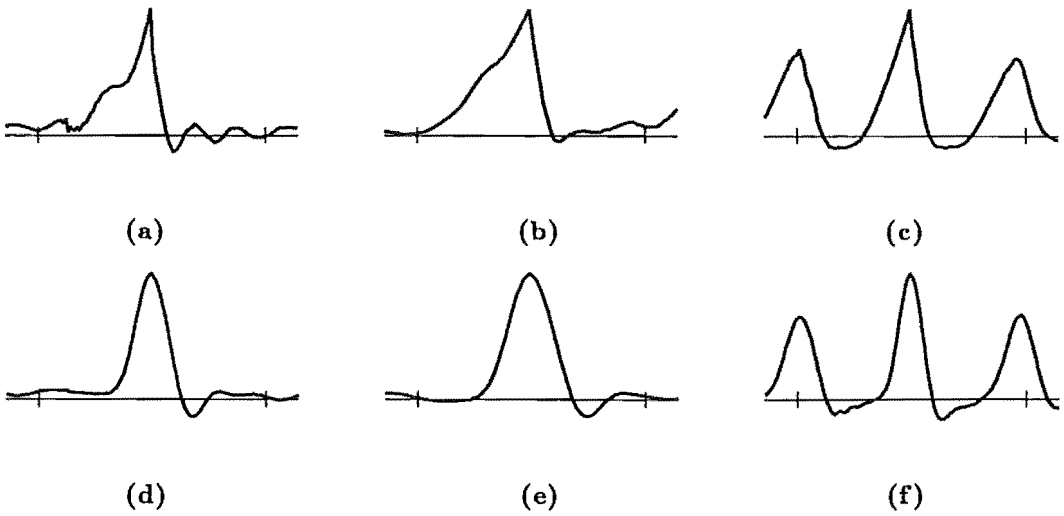


Figure 4.32. SAA signals relating to the synthetic utterances listed in Table 4.2. SAA was performed with $\tau_s = 12.8$ms and $\tau_p = 10$ms. **a:** AM-SAW, **b:** WM-SAW **c:** TF-SAW, **d:** AM-SIN, **e:** WM-SIN, and **f:** TF-SIN)

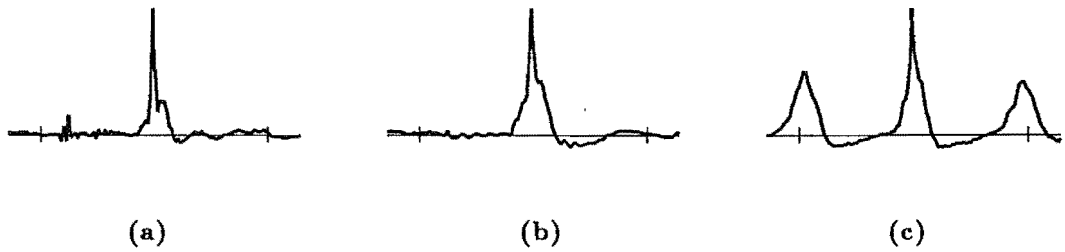


Figure 4.33. SAA signals obtained from the synthetic utterances **a:** AM-IMP, **b:** WM-IMP, and **c:** TF-IMP.

the invariant component of the filter, can be iteratively removed.

4.3.2.2 Synthetic speech generated from artificial filters

The results presented in §4.3.2.1 demonstrate the ability of SAA to extract estimates of the glottal pulse shape from (synthetic) speech signals. However, because the speech signals are synthesised from LPC coefficients taken from actual speech, the SAA signal contains contributions from the invariant component of the LPC filter. In order to evaluate the extent to which ghosting affects the shape of a SAA signal, I generated synthetic "vocal tract filter" signals $v_s(t)$ that had a negligible invariant component. These were used to generate synthetic speech-like signals by convolving them with fixed "glottal pulses". SAA signals were obtained from these synthetic signals and compared to the original pulses. Because of the negligible invariant components of the filter signals, any differences between the original pulse shapes and the SAA signals can be attributed to the effects of ghosting.

So that $v_s(t)$ has similar characteristics to an actual vocal tract filter response signal, I generate it from an ensemble of band-pass filters. Such a signal approximates

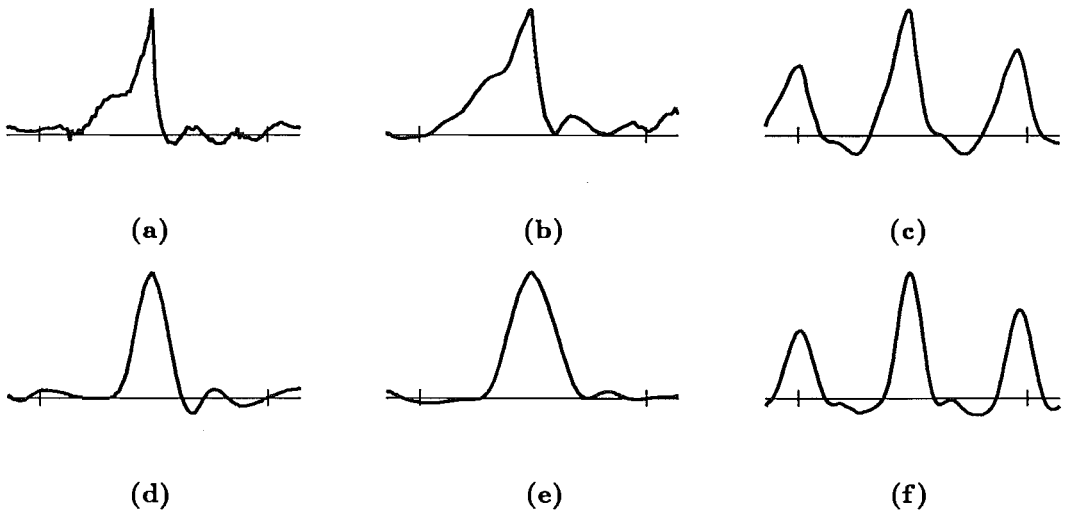


Figure 4.34. Estimates of the original pulse shapes obtained by deconvolving the "impulse-excited" SAA signals shown in Fig.4.33 from the appropriate SAA signals shown in Fig.4.32. A Wiener constant of 0.05 times the maximum spectral component in the filter was used in the deconvolution (see §1.3.2 for details on deconvolution techniques).

a single-formant model of speech, which, while inadequate to represent the phonetic information in speech (§2.1.4), does approximate the resonant characteristics of the vocal tract more than, for instance, white noise (which also has a negligible invariant component). In this model of speech, the vocal tract is represented as a single resonance having variable bandwidth and centre-frequency. $v_s(t)$ is composed of M concatenated segments, each of which is the impulse response $v_{sm}(t)$ of an ideal band-pass filter $V_{sm}(f)$. The bandwidths and centre-frequencies of these filters are determined so that the long term average spectrum of $v_s(t)$ is flat:

$$\sum_{m=1}^{M} |V_{sm}(f)| = K_s, \tag{4.28}$$

where M is the number of segments in $v_s(t)$ and K_s is an arbitrary (unspecified) constant. (4.28) can be satisfied in a straightforward way by constructing ideal band-pass filters $V_{sk}(f)$ which have spectral magnitudes given by (for positive frequencies)

$$V_{sk}(f) = \text{rect}\left(\frac{f - f_{BW}(k - \frac{1}{2})}{f_{BW}}\right), \quad k = 1 \ldots K \tag{4.29}$$

where $f_{BW} = f_H/K$ is the bandwidth of each of the K filters, f_H being the highest attainable frequency (i.e. half the sampling frequency in the discrete case). The impulse responses $v_{sk}(t)$ corresponding to the filters $V_{sk}(f)$ described by (4.29) are constructed from suitably frequency-shifted and time-compressed sinc functions.

In order to simulate the variation that occurs in the equivalent bandwidth of the vocal tract filter, I constructed several ensembles $\{v_{sk}(t)\}$ according to (4.29), with bandwidths f_{BW} ranging from 50Hz to 1000Hz in 50Hz steps. All such ensembles were then concatenated to form the synthetic time-varying vocal tract filter $v_s(t)$. In order to confirm that the $v_s(t)$ that is formed in this manner had a negligible invariant component, I computed the LTAS and SAA signal from $v_s(t)$. These are shown in Figs.4.35*a* and *b* respectively, and indeed confirm this conjecture.

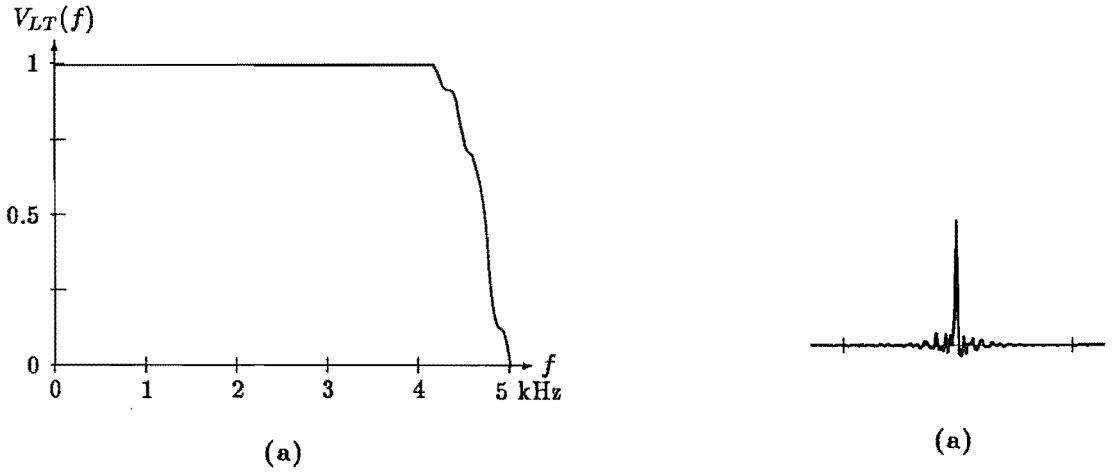


Figure 4.35. Demonstration that the invariant component of the synthetic filter $v_s(t)$ is negligible. **a:** LTAS of $v_s(t)$. **b:** SAA signal obtained from $v_s(t)$.

Synthetic speech-like utterances were constructed by convolving $v_s(t)$ with the synthetic glottal pulses shown in Figs.4.31*a* and *b*. Performing SAA on these utterances resulted in the SAA signals depicted in Figs.4.36*a* and *d* respectively. The SAA signal of the saw-tooth excited utterance has a saw-tooth shape except for a significant spike on top. By contrast, the SAA signal of the sine-squared excited utterance is much closer in shape to the original. Pre-emphasising each synthetic utterance before performing SAA leads to the SAA signals shown in Fig.4.36*b* and *e*, which, when integrated, appear as in Fig.4.36*c* and *f*. It is interesting that, when pre-emphasis is applied, the reconstruction of the saw-tooth is more faithful than that of the sine-squared pulse, while, without pre-emphasis, the reconstruction of the sine-squared pulse is more faithful than that of the saw-tooth.

The differences between the various SAA signals shown in Fig.4.36 can be understood by considering the different mechanisms involved. First-order differentiation of the saw-tooth pulse results in a signal with a large impulsive spike, corresponding to the trailing edge of the saw-tooth. The presence of this large spike means that SAA ghosting is greatly reduced. By contrast, first-order differentiation of the sine-squared pulse results in a "pulse" that is a complete period of a sinusoid. This contains two equal-magnitude peaks, which results in severe ghosting, since each peak is equally likely to be chosen in the SAA algorithm.

4.3.3 Refinement of true glottal pulse

As (4.13) explicitly defines, and §4.2.2 discusses in detail, the SAA signal contains contributions from the invariant components of both the glottal pulse and the vocal tract filter. If the purpose of SAA is to characterise the (average) glottal pulse alone, then some method for separating it from the vocal tract component is required. In this section I briefly describe a method of accomplishing this. This method, which is described more fully by Brieseman *et al.* (1987), iteratively produces an estimate of the glottal pulse from the SAA and LPC parameters of the speech utterance. Although this method of refining the glottal pulse estimate is yet to be comprehensively tested, it provides a starting point for examining the relationship between the glottal, vocal tract and ghost components of the SAA signal. Much further research is required to ascertain the properties of these relationships, and the methods and requisite conditions under

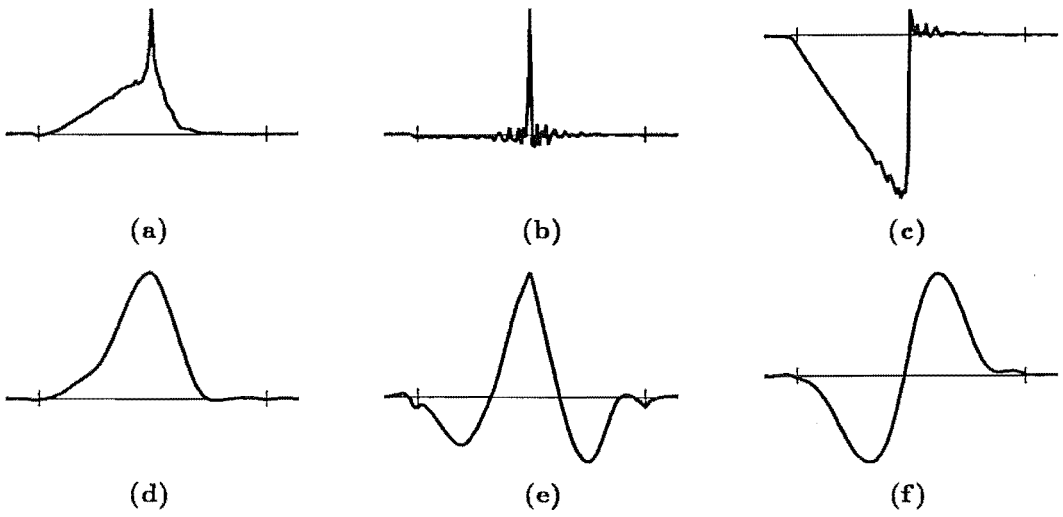


Figure 4.36. SAA signals obtained from the synthetic utterances generated by convolving $v_s(t)$ with **a:** saw-tooth and **d:** sine-squared pulses (see Fig.4.31). SAA signals of pre-emphasised utterances generated by **b:** saw-tooth and **e:** sine-squared excitations. Their first-order integrals appear in **c:** and **f:** respectively.

which they can be exploited to separate the different components of the SAA signal.

Linear prediction (LP) (§3.2) of a speech signal results in a time-varying all-pole filter that tends to match the formants or peaks of the (short-term) speech spectrum. These correspond to the resonances in the vocal tract, rather than to the glottal waveform characteristics (§3.4). Hence LP can be used to estimate the contribution of the vocal tract to the speech signal. As emphasised in §3.3.3 and §3.4.1, however, LP models the vocal tract accurately only if the analysis is performed when the glottis is closed. For the purposes of computational simplicity, I have not yet performed the processing in this way. Hence the LP filter coefficients are likely to be affected by the glottal excitation. This may be of little account because the inherent averaging in the subsequent SAA processing means that, if the effect of the glottal excitation on the LP filter coefficients varies enough when each LP analysis frame is aligned differently with the excitation, the average effect is reduced. In any case, these results give an indication of what can be expected with this technique.

By assuming that the LP filter characterises the vocal tract filter alone, it is reasonable to further assume that, if the excitation used to generate the synthetic speech is a train of "true" glottal pulses, the resulting SAA is necessarily identical to that obtained from the original speech utterance. Brieseman *et al.* (1987) propose an algorithm that iteratively refines an estimate of the true glottal pulse in such a way that the SAA signal obtained from the synthetic utterance converges to the SAA signal of the original speech utterance. This algorithm is described by the following sequence of steps:

1. Calculate $s_{sa}(t)$, the SAA signal of the original (voiced) speech utterance $s(t)$.
2. Abstract LP and pitch parameters from $s(t)$ in the usual way (§3.2 and §3.1.3).
3. Initialise the estimate of true glottal pulse as $g_p^{(1)}(t) = \delta(t)$.
4. Construct an excitation signal by replicating $g_p^{(\ell)}(t)$ according to the pitch infor-

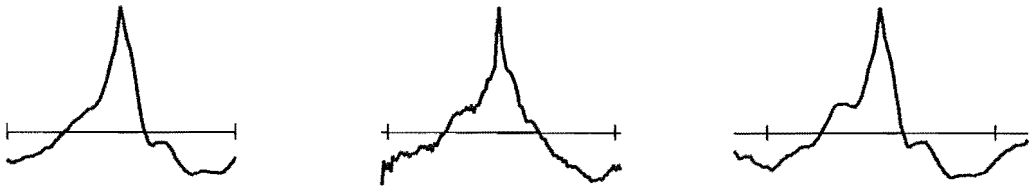


Figure 4.37. Iterative improvement of SAA signal to extract the glottal pulse component. **a:** $s_{sa}(t)$ of the utterance AM-RAIN2. **b:** $g_p(t)$, and **c:** $\sigma_{sa}(t)$ after 5 iterations.

mation obtained in step 2.

5. Compute a synthetic speech utterance $\sigma^{(\ell)}(t)$ from the excitation signal (step 4) and LP parameters (step 2).

6. Calculate $\sigma_{sa}^{(\ell)}(t)$, the SAA signal of $\sigma^{(\ell)}(t)$.

7. Compute the error $e^{(\ell)}(t)$ between the synthetic and original SAA signals:

$$e^{(\ell)}(t) = s_{sa}(t) - \sigma_{sa}^{(\ell)}(t), \quad -\tau_s/2 < t < \tau_s/2. \tag{4.30}$$

8. Update the estimate of the true glottal pulse

$$g_p^{(i+1)}(t) = g_p^{(\ell)}(t) + e^{(\ell)}(t), \quad -\tau_s/2 < t < \tau_s/2. \tag{4.31}$$

9. Repeat steps 4 through 9 until enough iterations have been performed or the average error $\eta^{(\ell)}$ is small enough:

$$\eta^{(\ell)} = \int_{-\tau_s/2}^{\tau_s/2} |e^{(\ell)}(t)| \mathrm{d}t < \eta_{thresh}. \tag{4.32}$$

The result of applying 10 iterations of the above algorithm to an utterance is illustrated in Fig.4.37, which also shows the "raw" SAA for the same utterance. As shown, the main peak in $g_p(t)$ is much "sharper" than that in $s_{sa}(t)$. Figs.4.38*a* and *b* show the effect of integrating $s_{sa}(t)$ and $g_p(t)$ respectively, in order to obtain the volume flow (§4.3.1). It appears from a comparison of Figs.4.38*a* and *b*, and by reference to the glottal flow signals shown in §4.3.1.1, that the iterative procedure results in a more accurate representation of the "true" glottal pulse (see also Brieseman *et al.* (1987) and Brieseman, 19XX).

One of the difficulties encountered in implementing the above algorithm is that the extremities of $g_p(t)$ are rarely equal to zero or even to each other. Hence step 4 tends to result in an excitation signal with gross discontinuities at the "joins" between each pulse. Furthermore, the synthetic speech arising from such an excitation, although it incorporates an SAA signal identical to the SAA of the original speech, is of much poorer perceptual quality. Further research is required to determine how the iteratively improved estimate of the glottal excitation can be sufficiently improved for it to be useful for speech analysis and synthesis (see §8.2.1).

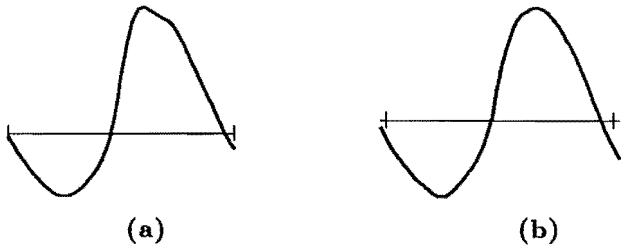


Figure 4.38. First-order integration of **a:** $s_{sa}(t)$, and **b:** $g_p(t)$ after 5 iterations, to illustrate the improved relationship that $g_p(t)$ has with the glottal flow (see Fig.4.27 in §4.3.1.1).

4.3.4 Discussion

From the results on synthetic utterances, it appears that SAA provides a good estimate of the invariant component of the glottal excitation. However, as shown by the differences between the SAA signals obtained from utterances formed out of entirely synthetic components, and those partly derived from actual speech signals, the SAA signal contains a significant contribution from the invariant component of the vocal tract filter response. The iterative improvement scheme described in §4.3.3 appears to remove some of the contribution from the vocal tract filter, but at the expense of much greater processing. The SAA signal also contains contributions caused by ghosting, which can be partially overcome by pre-emphasising the speech signal before obtaining the SAA signal.

The results of performing inverse filtering on different portions of an utterance indicate that the glottal excitation may vary significantly during an utterance (see §4.3.1.1). SAA, however, is only able to provide an estimate of the *average*, or *invariant*, excitation throughout the utterance. However, this estimate seems to be reasonably consistent with the average glottal excitation obtained via inverse filtering of the speech waveform (§4.3.1.1).

Because the SAA signal comprises both the average glottal excitation and the average (invariant) component of the vocal tract filter, it seems best to think of it as representing the long term characteristics of a person's voice, rather than just the glottal excitation. I treat it in this regard in Chapter 5, for the purposes of encoding speech at low data rates, and in §8.2.1.1, for the purposes of facilitating speaker recognition. Further interpretation of the SAA signal is presented in §5.4, together with a discussion of CLEAN processing of speech.

Chapter 5

Speech analysis by shift-and-add and CLEAN

The model of speech presented in Chapter 4 describes a speech signal as the convolution between invariant and time-varying components. Furthermore, the invariant part can be estimated by the blind deconvolution technique called shift-and-add (§4.2). All that is required to obtain the variant component is to deconvolve the shift-and-add component from the speech signal. Several approaches to deconvolving the shift-and-add signal from the speech signal are available (§1.3.2). This chapter describes in detail the subtractive deconvolution technique called "CLEAN", which was originally developed in the context of radio astronomy (§5.1.1). The advantages and disadvantages of Wiener filtering, which is an alternative deconvolution technique, are discussed in §5.1.3.

In §5.1 I briefly describe the astronomical background of CLEAN, and thereby introduce the terminology used by Högbom (1974) to describe the technique. §5.1 also refers to the several other areas to which CLEAN has been applied, and discusses its relationship to Wiener filtering. The details of how the algorithm is applied to speech signals are described in §5.2, while §5.3 presents the results obtained by employing SAA and CLEAN in a low data rate speech encoding scheme. Finally, a detailed discussion of various ways that the SAA/CLEAN approach to speech analysis can be interpreted appears in §5.4

5.1 Background

5.1.1 Astronomical origins

CLEAN is a method of subtractive deconvolution that arose in the context of synthesis radio astronomy (Högbom, 1974). §5.1.1.1 briefly describes the technique, employed in radio astronomy, of "synthesising" a very large telescope from several small, widely spaced, telescopes. §5.1.1.2 then describes the CLEAN algorithm as presented by Högbom (1974). Hogböm's terminology for describing CLEAN is also introduced in §5.1.1.2.

5.1.1.1 Synthesis telescopes in radio astronomy

In radio astronomy, two or more antennas are instrumented to measure the magnitude and phase of the incoming radiation, by means of interferometry (Napier *et al.*, 1983; Thompson *et al.*, 1986). Each measurement provides data for a single point in the aperture of a very large "synthetic" telescope (Christiansen and Högbom, 1969). The

aperture plane is related by the (two-dimensional) Fourier transform to the *image plane*, in which the image of the heavens is revealed. As stated in §1.2.5.4, the resolution in one domain is inversely proportional to the extent of the signal in the other domain. Hence synthesis radio astronomy is a method of obtaining far greater resolution than that possible from a single antenna. However, because the aperture plane is only sampled at a few points, the synthetic image is severely *blurred*. The blurring can be described by considering the measured aperture distribution $R(u,v)$ (where u and v are the two-dimensional Fourier coordinates) to be the product of the true aperture distribution $T(u,v)$ and a sampling function $G(u,v)$ that is unity at discrete points and zero elsewhere. By application of the convolution theorem, the resulting image $r(x,y)$ is given as (Högbom, 1974)

$$r(x,y) = \mathcal{F}^{-1}\{T(u,v).G(u,v)\} \tag{5.1}$$
$$= t(x,y) \odot g(x,y) \tag{5.2}$$

where $t(x,y)$ is the true image and $g(x,y)$ is the blurring function.

$g(x,y)$ in (5.1) typically exhibits large sidelobes that may extend across the whole extent of $r(x,y)$. Because $G(u,v)$ consists of only a few non-zero components, it is not realistic to employ Wiener filtering to deconvolve it from the measured image. In fact, deconvolution of $G(u,v)$ is effectively the same as interpolating between the sampled values of $T(u,v)$. In general this is of course not possible (except from specialised points of view such as underly, for instance, the maximum entropy principle in image processing contexts, cf. Narayan and Nityananda, 1986; Cornwell, 1988) without taking into account some *a priori* information about the distribution of astronomical objects (Högbom, 1974). An example of such information is that the true image consists primarily of only a few point sources distributed throughout the image (Schwartz, 1978).

5.1.1.2 The CLEAN algorithm

The reasoning that leads to the technique of CLEAN follows directly from the point made in the last paragraph of §5.1.1.1. Rephrased, it states that $r(x,y)$, hereafter called the *dirty map*, is the result of a convolution between a few point sources (which collectively comprise $t(x,y)$, the *CLEAN map*) and $g(x,y)$, termed the *dirty beam*. This convolution can be expressed as

$$r(x,y) = \sum_{i=1}^{N} \beta_i g(x - x_i, y - y_i) \tag{5.3}$$

where N is the number of discrete points in the image, the i^{th} of which is located at position (x_i, y_i) and has amplitude β_i. The pulse positions and amplitudes can be estimated iteratively by the *CLEAN* algorithm (Högbom, 1974):

1. The peak of maximum magnitude in the dirty map is located.

2. The dirty beam, weighted both by the amplitude of the peak identified in step 1 and a suitable factor (called the *loop gain*), is centred on the peak position and subtracted from the dirty map.

3. Steps 1 and 2 are repeated, with the dirty map replaced by the remainder from the previous iteration, until the maximum value is deemed to be no longer significant.

4. All the pulses identified in step 1 are added together to form the CLEAN map.

5. The restored image is obtained by convolving the CLEAN map with the CLEAN beam, which is the ideal point spread function corresponding to a completely filled, uniform aperture.

Readers who wish to find out more about the radio-astronomical applications of CLEAN are referred to Högbom (1974) for the original treatment, Schwartz (1978) for an in-depth mathematical-statistical analysis of the algorithm, and Chapter 11 of Thompson *et al.* (1986) for an up to date general overview of the subject.

5.1.2 Application of CLEAN to other deconvolution problems

CLEAN has also been applied to deconvolution problems in several diverse fields (Bates *et al.*, 1982a). Millane and Bates (1982) invoke CLEAN to deblur transmission line reflections in order to improve their method of time domain reflectometry. The true signal in this application consists of discrete impulses corresponding to discontinuities in the transmission line. However, the echoes that are actually recorded are blurred, so it is desirable to remove this blurring function before attempting to find any particular "inverse solution". CLEAN is appropriate for performing this "deblurring" because it produces an impulsive signal. However, the loop gain must be very small to ensure that the impulse positions are located accurately (Bates *et al.*, 1982a).

Another application where CLEAN has been successful is the deblurring of biochemical analysis data (Bates, 1981; Bates *et al.*, 1982a). Such data tend to exhibit position-variant blurring, so that ordinary multiplicative deconvolution (§1.3.2) is inappropriate. CLEAN can be invoked to resolve the exact positions of several peaks (for example in a record of chromatographic data) that otherwise tend to be blurred together (Bates *et al.*, 1982a).

5.1.3 Comparison with Wiener filtering

CLEAN is termed a *subtractive* deconvolution technique, while Wiener filtering (§1.3.2) is called *multiplicative* deconvolution (Bates and McDonnell, 1986, Chapter III). In this section I discuss the relationship between these two types of deconvolution.

Wiener filtering consists of multiplying the Fourier transform of the blurred signal with an *inverse filter* formed from an estimate of the blurring function (Note that the filter can also be applied equivalently as a convolution in the time domain). The deconvolution is thus performed as a "single" operation. By contrast, CLEAN consists of iteratively subtracting scaled copies of the blurring function from the blurred signal. It thus does not involve any Fourier transformation, but nevertheless requires greater computational effort (Bates *et al.*, 1984).

Because CLEAN operates by means of repeated subtractions in the time domain, it is straightforward to allow the blurring function to be time-variant. Wiener filtering can only deconvolve a time-variant blurring function if the signal can be usefully partitioned into segments over each of which the blurring function is invariant. This is the rationale behind many of the inverse filtering techniques invoked in speech analysis (which are themselves species of deconvolution cf. §3.4.1; §3.5.2).

Another situation in which multiplicative deconvolution is inappropriate arises when the data contain "missing values". Indeed, the CLEANing of radio-astronomical images is motivated by large areas of Fourier space being zero-valued (§5.1.1.1). CLEAN is also more reliable when the data are truncated in the time domain (Bates *et al.*, 1982b).

Even when large amounts of Fourier data are "missing", it is straightforward to CLEAN the data with the actual blurring function, which effectively interpolates

the missing values in the Fourier domain (cf. Bates *et al.*, 1984; Thompson *et al.*, 1986, §11.1).The CLEAN signal is then re-convolved with the "ideal" blurring function (i.e. step 5 of the procedure introduced in §5.1.1.2). Wiener filtering can then be invoked to complete the deconvolution. In this way the computational efficiency of Wiener filtering can be combined with the ability of CLEAN to cope with missing data values (Bates *et al.*, 1984).

Because CLEAN produces an impulsive CLEAN signal, regardless of the form of the original, undistorted, signal, intricate precautions have to be taken when attempting to reconstruct "smooth" signals (Bates *et al.*, 1982b; Cornwell, 1983).

5.2 CLEAN processing of speech signals

The purpose of applying CLEAN to speech is to deconvolve out its invariant characteristics, as represented by the SAA signal (§4.2). The resulting "CLEAN signal" thus represents the dynamic or time-varying aspects of the speech signal. As described in §8.2.1.2, it is possible to perform this by means of standard Wiener filtering techniques. However, CLEAN has several advantages over Wiener filtering that make its use in this application much more attractive. The main advantage is that it produces a signal that consists of relatively few non-zero "pulses". As shown by the results presented in §5.3, this means that the CLEAN signal can be directly encoded at low data rates, without the need for further processing (the usefulness of CLEAN as a pre-processor for LPC schemes is introduced in §8.2.2.4). The second advantage of CLEAN over Wiener filtering is that it is not troubled by the smallness of the higher frequency components in the spectrum of the SAA signal (Högbom, 1974).

This section discusses the application of the CLEAN algorithm to speech signals. §5.2.1 introduces appropriate terminology. The characteristics of speech signals that must be considered in order to successfully employ the CLEAN algorithm are discussed in §5.2.2, while a modified algorithm is presented in §5.2.3. Details of the latter algorithm's implementation are discussed in §5.2.5. In §5.2.6, a technique of improving the estimates of the pulse amplitudes is described.

5.2.1 Application of the CLEAN algorithm to speech

The model of speech introduced in §4.2 describes a speech signal in terms of the invariant (from one pitch interval to another) and variant components of the glottal excitation and vocal tract filter. Furthermore, the invariant component can be simply and efficiently estimated by shift-and-add (§4.3). Although methods for separating the invariant glottal and vocal tract components are available (§4.3.3), they are not necessary for the types of processing described in this chapter. Following this approach, a speech utterance $s(t)$ is represented as

$$s(t) = s^i(t) \odot s^v(t) + s^c(t) \tag{5.4}$$

where $s^i(t)$ is the invariant part of the speech signal, $s^v(t)$ is the variant component, and $s^c(t)$ is the "contamination".

It is convenient to employ Högbom's terminology when discussing the CLEAN algorithm. The speech signal $s(t)$ then becomes the "dirty signal", while the term "CLEAN signal" is applied to the deconvolved signal $s^v(t)$. I refrain from calling the SAA signal $s^i(t)$ the "dirty beam", because I do not have an associated CLEAN beam as is appropriate in radio astronomy. Instead, I refer to it as the CLEAN kernel, or SAA signal, whichever is the more appropriate at the time. Although this graphic terminology suggests several interesting applications for the technique of "CLEANing" speech,

feats such as automatic censoring of movie sound tracks are beyond the algorithm's abilities!

5.2.2 Speech characteristics relevant to CLEAN processing

Several characteristics possessed by speech signals have forced me to alter the basic CLEAN algorithm introduced in §5.1.1.2. The SAA signal, obtained by the process described in §4.2.4, is only valid for the voiced sections of a speech utterance. Two methods of accounting for the unvoiced sections are available. The first is to employ standard VUV techniques (§3.1.2) to section the speech signal into voiced and unvoiced parts. These can either be processed separately, or a different kernel can be employed for the voiced and unvoiced sections, thus taking advantage of the ability of CLEAN to utilise a time-variant kernel (Bates *et al.*, 1982c) I have found that a better approach is, however, to separate the speech into two sub-bands (cf. §3.5.1.2), the lower of which contains most of the energy of the voiced speech sections, while the high frequency band contains most of the unvoiced energy. Each sub-band can then be processed separately by SAA and CLEAN. §5.2.5.5 and §5.4.3.3 discuss the details and implications respectively of performing CLEAN on sub-bands of the speech signal.

Another difference between speech and astronomical signals is that speech signal is of arbitrarily long duration, while most astronomical images are of fixed and quite limited extent. Many applications, however, require that speech be processed in *segments* of short duration, so that only a small delay is introduced by the processing. For example, low data rate encoding schemes must analyse the incoming speech in "real time", with a delay between the input and output of the system of several tens of milli-seconds at the most. For this reason, the CLEAN algorithm must be modified to process short segments of speech in a sequential fashion, rather than the whole signal at once as described in §5.1.1.2. The details of how the speech is segmented are described in §5.2.5.3.

A further characteristic of speech signals is that they vary a great deal from one segment to another. Although $s^i(t)$ is presumed to represent the component of the speech signal that remains the same for all segments, it is defined by an average over many segments. Any one particular segment, however, may comprise components that are quite different from the average, without significantly affecting that average. Hence, the SAA signal may not realistically represent one component of a convolution making up that segment (if the "contamination" for that segment is relatively high). An extreme example of this is of course an unvoiced segment, as intimated in the first paragraph of this section. An unvoiced sound cannot be realistically modelled as a convolution between the SAA signal obtained from voiced segments of speech and some other component, even though the SAA signal obtained from both voiced and unvoiced segments of speech is not very different from that obtained from voiced segments only (§4.2.4.5). Further discussion of this point appears in §5.4.3.2.

Even though a particular speech segment may not be modelled "realistically" by a convolution between the SAA signal and some other component, $s^i(t)$ can be deconvolved from it to leave a "variant" component $s^v_m(t)$. As indicated in the next two paragraphs, it may not matter that $s^v_m(t)$ is not physically realistic.

The original justification for the development of CLEAN was that many astronomical images are dominated by only a few isolated "spikes". By performing CLEAN with a low enough loop gain, the spikes in the resulting CLEAN map are therefore justifiably assumed to accurately represent the true image. The validity of this assumption has been confirmed by experiments with computer generated images (cf. Bates *et al.*, 1982b; Bates *et al.*, 1984). When reconstructing signals that are continuous, however, the resulting CLEAN map still has a "spiky" appearance (compare Fig.5.1*e* with

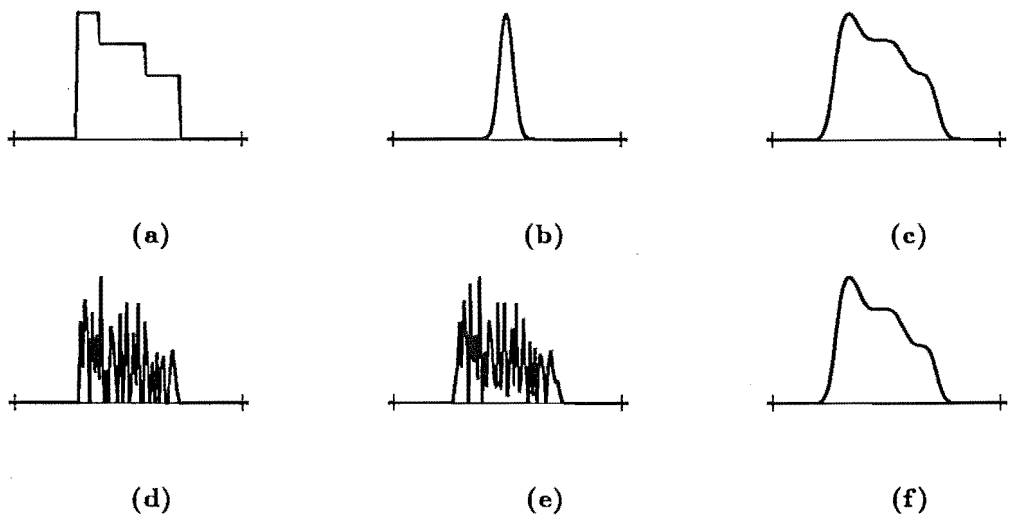


Figure 5.1. Application of CLEAN to a smooth (non-spiky) signal. **a:** Original signal. **b:** Blurring function (Gaussian). **c:** Blurred version of the signal shown in *a* (the dirty signal). **d:** CLEAN signal after 50 iterations with a loop gain of 0.2, showing its "spiky" nature. **e:** CLEAN signal after 200 iterations. **f:** The "synthetic" blurred signal formed by convolving the CLEAN signal shown in *d* with the blurring function shown in *b*.

Fig.5.1*a*; also see Bates *et al.*, 1984). Nevertheless, this spiky signal is still a possible solution to the deconvolution problem, as evidenced by the close similarity between the blurred version of the original signal (Fig.5.1*c*) and the reconstruction formed from the (CLEANed) spiky signal (Fig.5.1*f*). The small difference between Figs.5.1*c* and *f* (even though the CLEANed signal shown in Fig.5.1*e* bears little resemblence to the "original" signal shown in Fig.5.1*a*) implies that a variety of different signal pairs $s^v(t)$ and $s^c(t)$ are able to adequately satisfy the speech model (5.4).

For the purpose of representing the speech signal for later resynthesis, the exact form of $s^v(t)$ does not matter, as long as $s(t)$ can be recovered from it. In fact, for the method of low data rate encoding described in §5.3, the sparse nature of the CLEAN signal is a great asset.

Despite the adequacy of a "spiky" CLEAN signal as a representation of $s^v(t)$, in the sense that $s^v(t)$ can be faithfully reconstructed from it, the question remains as to whether it represents any "real-world" speech component (such as what the excitation signal and filter are supposed to represent in LP analysis). This question is discussed further in §5.4.1.

5.2.3 CLEAN algorithm for speech signals

The modified CLEAN algorithm for a segment of speech is defined by the following sequence of steps. Initially, $r_1(t)$ is set equal to $s(t)$ and $v_0(t) = 0$, for $0 < t < \tau^{\text{seg}}$, where τ^{seg} is the duration of the segment (see §5.2.5.3 for details on segmentation). The CLEAN kernel $g(t)$ is set to $s_{sa}(t)$ or a modified form of $s_{sa}(t)$ (see §5.2.5.1). Thereafter, for each iteration labelled by j:

1. The position of the new pulse p_j is located at the instant

$$p_j = \arg\max_t |r_j(t)|, \tag{5.5}$$

2. The amplitude of the new pulse v_j is given by

$$v_j = \gamma r_j(p_j)/g(0) \tag{5.6}$$

where γ is the loop gain (see §5.2.5.2).

3. The CLEAN signal is updated:

$$v_j(t) = v_{j-1}(t) + v_j\delta(t - p_j), \quad 0 < t < \tau^{\text{seg}}. \tag{5.7}$$

4. If the pulse position p_j signifies a new distinct pulse in $v(t)$ (i.e. $v_{j-1}(p_j) = 0$), the pulse counter N_p is incremented.

5. The dirty signal is reduced:

$$r_{j+1}(t) = r_j(t) - v_j g(t - p_j), \quad p_j + \tau_g^- < t < p_j + \tau_g^+ \tag{5.8}$$

where $\tau_g^- < t < \tau_g^+$ is the interval over which $g(t)$ is non-zero when the maximum magnitude in $g(t)$ occurs at $t = 0$.

6. Steps 1–5 are repeated until either $|v_j| < \eta_{\text{cl}}$, $N_p = P_{\max}$, or $j \geq J_{\max}$ (see §5.2.5.4 for details on each of these threshold parameters).

It is appropriate here to introduce a special notation for the non-zero samples of the CLEAN signal (i.e. the samples of $v(t) \neq 0$). I write this set of non-zero samples as $\{v_i; p_i, i = 1, 2 \ldots P_{\max}\}$, which I call the sequence of CLEAN pulses.

5.2.4 Reconstructing speech from the CLEAN signal

Speech can be reconstructed from the CLEAN signal $v(t)$ simply by convolving it with the kernel signal $g(t)$. The convolution can be computed very efficiently if one takes appropriate advantage of the large number of zero-valued samples in $v(t)$. The following sequence of steps represents one way of doing this:

1. Initialise $\hat{s}(t) = 0$, $0 < t < \tau^{\text{seg}}$

2. Repeat the following step for each of the non-zero samples in $v(t)$, the k^{th} of which is of amplitude v_k and position p_k.

3. Add the kernel corresponding to the k^{th} pulse to $\hat{s}(t)$:

$$\hat{s}(t + p_k) = \hat{s}(t + p_k) + v_k g(t), \quad \tau_g^- < t < \tau_g^+. \tag{5.9}$$

5.2.5 Implementation details and considerations

In this section I provide details on the implementation of the CLEAN algorithm presented in §5.2.3. §5.2.5.1 explains how the SAA signal is modified so that it can be used as the CLEAN kernel, while §5.2.5.2 describes the effect that the loop gain factor γ has on the operation of the CLEAN algorithm. §5.2.5.3 discusses the need to segment the speech utterance before performing CLEAN, and describes how this can best be done. §5.2.5.4 discusses the various conditions that can be invoked to terminate the CLEAN algorithm, while §5.2.5.5 describes how CLEAN can be applied to the unvoiced sections of speech. Finally, §5.2.5.6 describes the improvement in performance that is obtained when the speech signal is pre-emphasised before performing SAA and CLEAN.

Parameter	Default value	Description
γ	0.7	Loop gain
τ_k^{seg}	25ms	Segment duration
ΔT	15ms	Segment spacing
η_{cl_g}	$0.01 \max\{\lvert s(t)\rvert\}$	Global threshold
η_{cl_k}	$0.05 \max\{\lvert s_k(t)\rvert\}$	Local threshold
$P_{\max}$	variable	Maximum number of pulses per segment
$J_{\max}$	$2P_{\max}$	Maximum number of iterations per segment

Table 5.1. Default values for the parameters in the CLEAN algorithm. See §5.2.5.1 through §5.2.5.6 for details

5.2.5.1 Necessary modifications to the SAA signal

My implementation of the SAA procedure described in Chapter 4 produces a signal that is of fixed duration (12.8ms) for all utterances, no matter what the average pitch of the utterance is. While this is of no concern for the purposes of Chapter 4, it is necessary to modify the SAA signal before it can be used in the CLEAN algorithm described in this chapter. The SAA signal is, firstly, usually longer than the average pitch period of the utterance and, secondly, its end-points are seldom of zero amplitude. Consequently, discontinuities are introduced into the dirty signal when the SAA signal is subtracted from it at step 5 of the CLEAN algorithm (§5.2.3). In this section I discuss several strategies for modifying the SAA signal so that it is better suited for use in CLEAN processing.

Extracting a portion of the SAA signal of duration equal to the average pitch period of the utterance, as is illustrated in Fig.5.2*a*, does not in general result in a signal whose end-points are of zero amplitude. It is thus necessary to subtract an offset equal to the average value of the end-points from the SAA signal.

Fig.5.2 illustrates several ways in which the SAA signal is modified for use in the CLEAN algorithm. The "raw" SAA signal $s_{sa}(t)$ obtained from the utterance AM-RAIN1 is shown in Fig.5.2*a*. Instants τ_g^- and τ_g^+ are identified at locations having approximately the same values and slopes of $s_{sa}(t)$. An offset equal to $(s_{sa}(\tau_g^-) + s_{sa}(\tau_g^+))/2$ is subtracted from $s_{sa}(t)$ and the result, for $\tau_g^- < t < \tau_g^+$, denoted by $g(t)$. The signal $g(t)$ is shown in Fig.5.2*b*.

Because the addition of a d.c. offset to the SAA signal is likely to affect the operation of CLEAN (because it is a method of subtractive deconvolution), I have investigated several other methods of modifying the SAA signal so that its end-points are of zero amplitude. One approach is to identify the instants τ_g^- and τ_g^+ as illustrated in Fig.5.2*a*, but instead of subtracting an offset, modify $s_{sa}(t)$ outside the interval $\tau_g^- < t < \tau_g^+$ so that it goes smoothly to zero. The resulting $g(t)$, which is longer than τ_g because of the "edge-extension", is shown in Fig.5.2*c*. This technique has been successful in countering the effects of truncation when reconstructing a wide variety of classes of signals and images (Bates and McDonnell, 1986, §15). Another approach is to multiply the SAA signal by a tapering window (cf. §1.3.1.1) so that its values at τ_g^- and τ_g^+ are zero. A final approach to obtaining $g(t)$ from $s_{sa}(t)$ is to position the end-points at instants where a zero-crossing occurs in $s_{sa}(t)$, as shown in Fig.5.2*d*. Unlike in the other methods outlined in this and the previous paragraph, the SAA signal values are

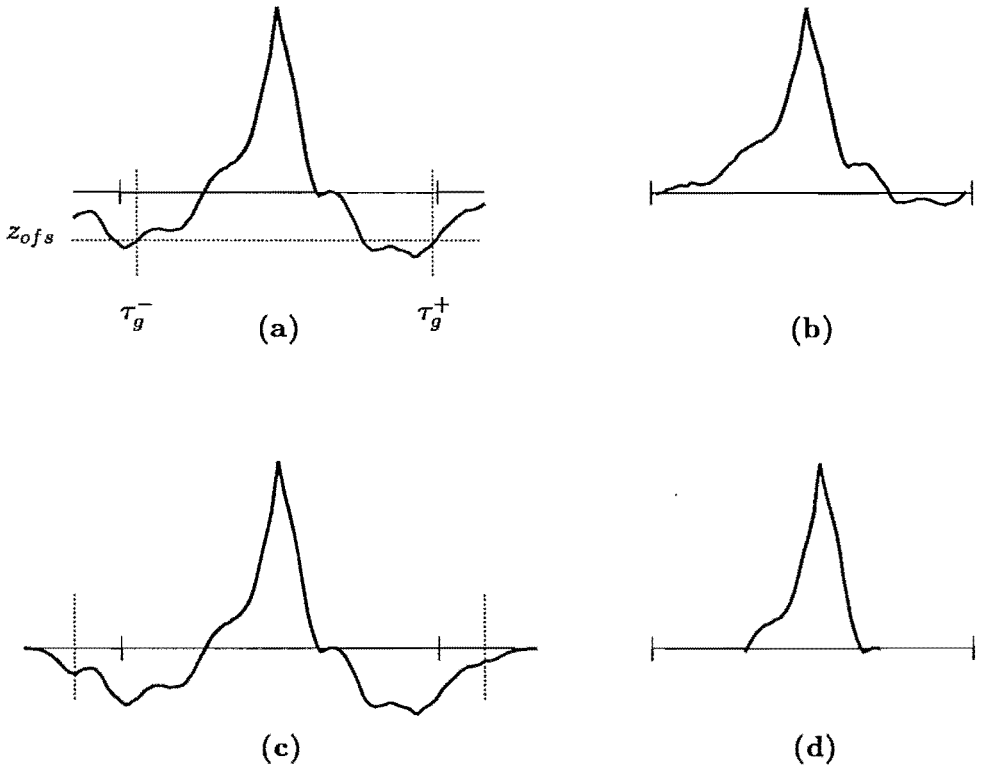


Figure 5.2. Illustration of ways in which the SAA signal $s_{sa}(t)$ is modified to form the CLEAN kernel $g(t)$. **a:** Raw SAA signal, showing the positions of the end-points τ_g^- and τ_g^+ respectively. **b:** CLEAN kernel obtained by subtracting an offset from the SAA signal, such that the values of $g(t)$ at the end-points shown in *a* are zero. **c:** "Edge-extending" $s_{sa}(t)$ with a smooth Gaussian roll-off from each of the end-points identified in *a* to zero amplitude. **d:** Positioning the end-points of $g(t)$ at zero-crossing instants in $s_{sa}(t)$.

not actually changed in this approach.

Experiments with each of the methods described above convinced me that the simple approach of subtracting a d.c. offset so that the end-points $s_{sa}(\tau_g^-)$ and $s_{sa}(\tau_g^+)$ are nearly zero leads to the best performance as far as subsequent CLEAN processing is concerned. All the results presented in this chapter are therefore generated with SAA signals modified in this manner.

5.2.5.2 Choice of gain parameter

This section discusses how changing the value of the loop gain factor γ affects the operation and results of the CLEAN algorithm described in §5.2.3. It is first necessary to introduce some fresh terminology. The CLEAN algorithm produces a CLEAN signal and a *residual* dirty signal. The CLEAN signal represents the convolutional component of the speech signal, while the residual signal represents the additive contamination (§5.2.1). In this and subsequent sections I often refer to the *level* of the dirty signal. By this I mean the average (squared) magnitude of the dirty signal compared to the original speech signal. The level of the dirty signal r_{lev} is defined by

$$r_{\text{lev}} = \langle |r(t)|^2 \rangle_t / \langle |s(t)|^2 \rangle_t, \tag{5.10}$$

which is seen to be related to the signal-to-noise ratio SNR by the expression

$$\text{SNR} = -10 \log_{10}(r_{\text{lev}}). \tag{5.11}$$

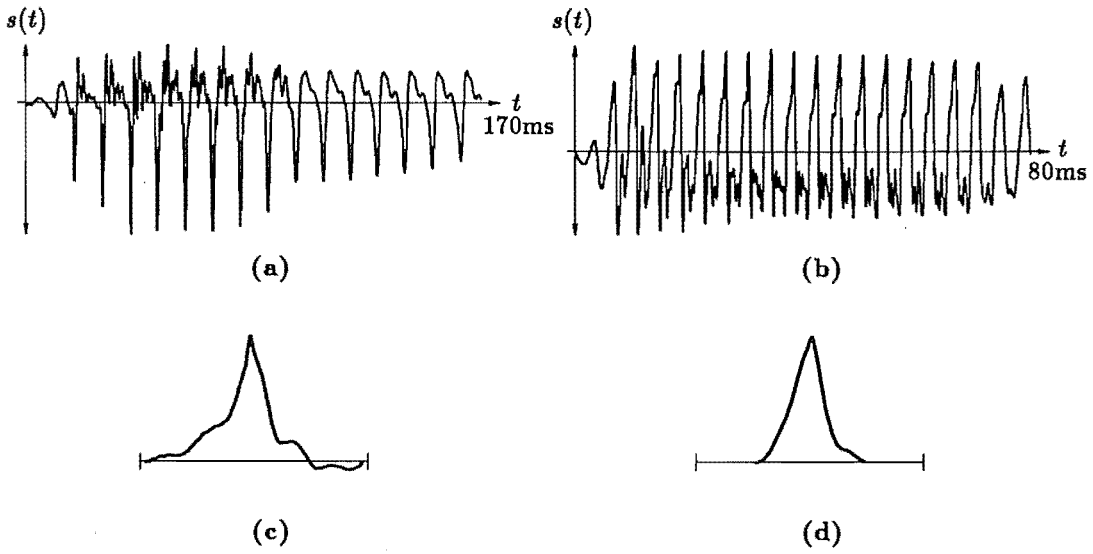


Figure 5.3. Segments of speech invoked to illustrate the operation of the CLEAN algorithm. **a:** Segment of the utterance AM-RAIN1. **b:** Segment of the utterance TF-RAIN1. **c:** Modified SAA signal from the utterance AM-RAIN1. **d:** Modified SAA signal from the utterance TF-RAIN1

When describing the convergence of the CLEAN algorithm, I find it convenient to employ the terminology r_{lev}. When discussing the CLEAN algorithm in terms of low data rate speech encoding, however, the more conventional SNR is invoked.

The value of the loop gain γ strongly influences the rate of convergence of the CLEAN algorithm. A smaller value for γ means that a larger number of iterations are required to CLEAN a dirty signal down to any given level. This is because the dirty signal is reduced by only small "nibbles" at each iteration, instead of large "bites" when a larger value is used. Fig.5.4 shows the SNR versus the number of iterations when the segments of speech shown in Fig.5.3*a* and *b* are CLEANed with the SAA signals shown in Fig.5.3*c* and *d* respectively. Results are shown for several values of γ ranging from 0.1 to 1.0.

Figs.5.5 to 5.7 show the CLEAN residual, and reconstructed signals that arise when the segments of speech shown in Figs.5.3*a* and *b* are CLEANed with various values of γ ranging from 0.1 to 1.0. The dirty signals shown in both Figs.5.5 to 5.7 have been reduced to a similar level, but the forms of the CLEAN signals are very different. In particular, the number of non-zero CLEAN pulses is very much greater when γ is smaller. Fig.5.8 shows curves of SNR versus the number of non-zero samples in the CLEAN signals obtained as described above.

The value of the CLEAN loop gain γ can be viewed as an estimate of the level of confidence in how well the amplitude of the current maximum peak in the dirty signal represents the amplitude of the CLEAN pulse at that instant. For instance, if one knows that the CLEAN signal consists of well separated pulses, one can assume that a peak in the dirty signal is composed of a single copy of the kernel, positioned on that peak. A reasonable value for γ is therefore unity. On the other hand, if the dirty signal is composed of many overlapping copies of the CLEAN kernel, the maximum peak is likely to be composed of the superposition of many overlapping copies of the CLEAN kernel. Hence the amplitude of the peak is likely to be appreciably different from the amplitude of the CLEAN pulse at that position, implying that a low value for γ is advisable. The results presented here suggest that γ should not be too small,

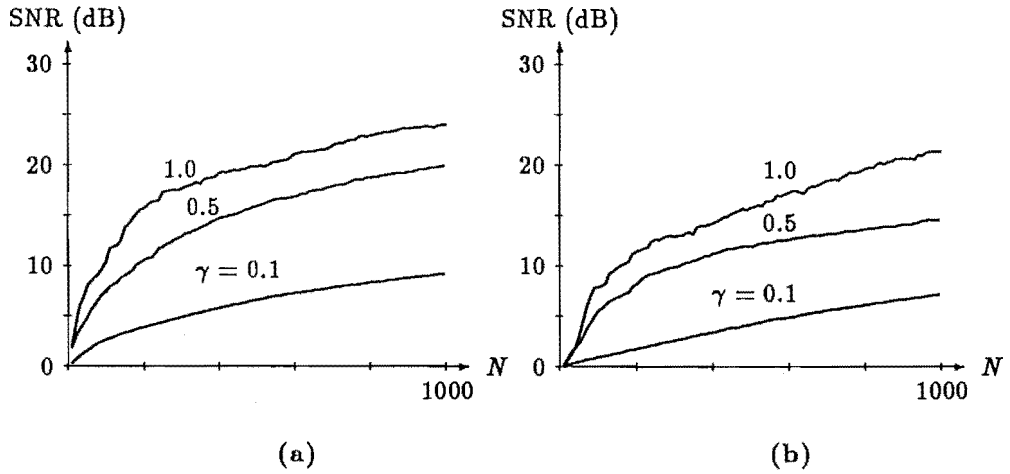


Figure 5.4. SNR as a function of the number N of CLEAN iterations applied to the speech segments shown in Fig.5.3. **a:** pertains to the segment shown in Fig.5.3*a* while **b:** pertains to the segment shown in Fig.5.3*b*. Curves are shown for values of loop gain γ equal to 0.1, 0.5, and 1.0.

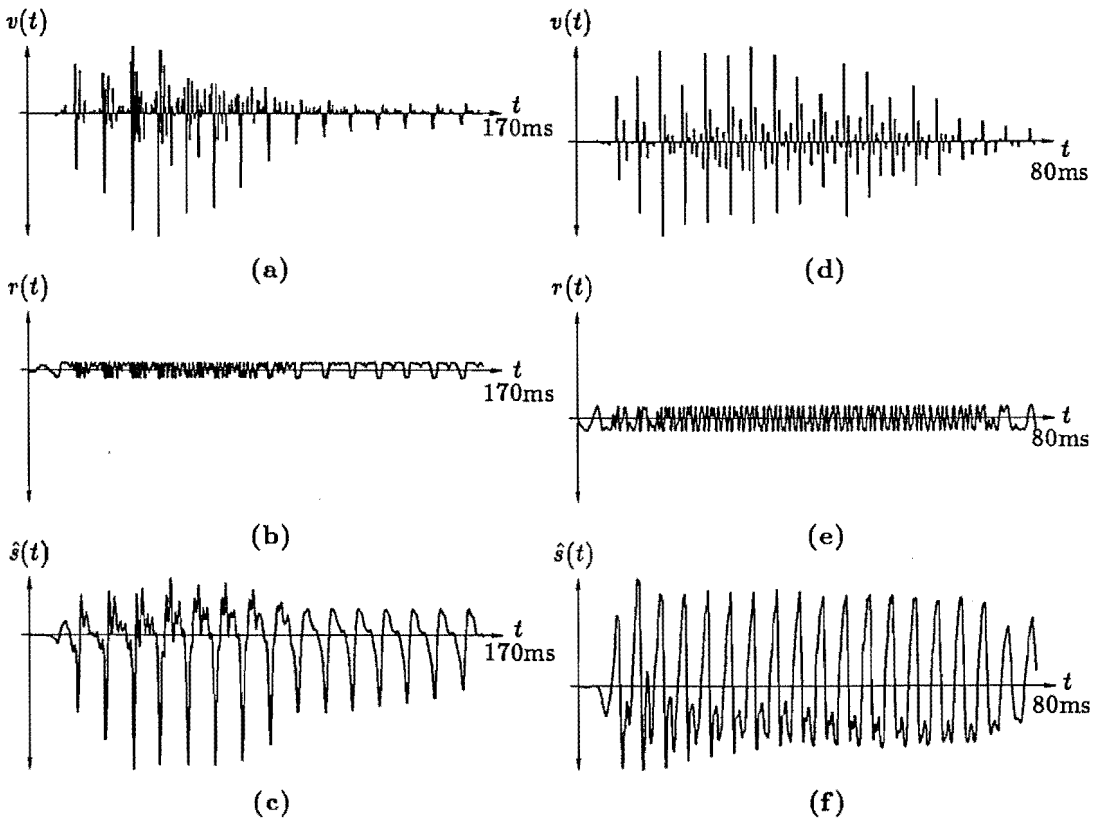


Figure 5.5. The results of applying CLEAN with a loop gain γ of 0.1, to the speech segment shown in Figs.5.3*a* and *b*. **a:** CLEAN, **b:** residual, and **c:** reconstructed signals obtained from the segment shown in Fig.5.3*a*. **d:** CLEAN, **e:** residual, and **f:** reconstructed signals obtained from the segment shown in Fig.5.3*b*. 2800 CLEAN iterations are performed to produce the CLEAN signal shown in *a*, while 6500 are required to produce the CLEAN signal shown in *d*. The resulting SNR is 15dB in each case. The number of non-zero "pulses" in the CLEAN signals shown in *a* and *d* are equivalent to 2700 and 3100 pulses per second respectively.

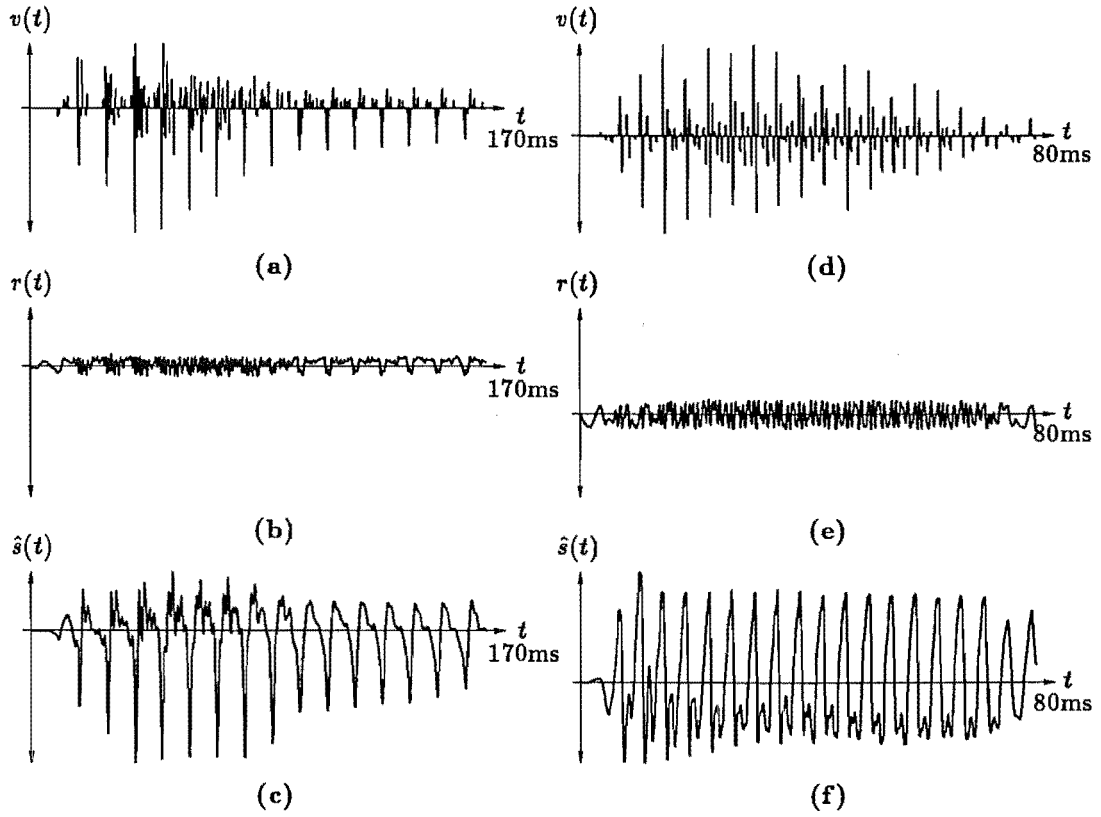


Figure 5.6. The results of applying CLEAN with a loop gain γ of 0.5, to the speech segment shown in Figs.5.3*a* and *b*. **a:** — **f:** are the same as in Fig.5.5. 430 and 1000 CLEAN iterations respectively are performed for **a:** — **c:** and **d:** — **f:**, resulting in SNR = 15dB in each case. The CLEAN signals shown in *a* and *d* contain the equivalent of 1580 and 2760 pulses per second respectively.

both to avoid the stagnation evident in Fig.5.4, and to reduce the number of non-zero samples in the CLEAN signal. Conversely, it should not be large, or else instability may result. For the remainder of this chapter, I employ a value of $\gamma = 0.7$ unless otherwise specified.

5.2.5.3 Segmentation considerations

Because a speech utterance is a signal of arbitrarily long duration, it is necessary to divide it into short *segments* of convenient duration before performing CLEAN. Each segment is processed by the CLEAN algorithm described in §5.2.3, with the total CLEAN signal being formed by concatenating the individual CLEAN signals from each segment. However, care must be taken to avoid "segmentation edge effects" which can occur when an utterance is segmented arbitrarily. In this section I describe these effects and how they can be minimised.

Before discussing the details of the segmentation edge effects, it is useful to introduce some pertinent terminology. A speech utterance $s(t)$ is divided into K segments $s_k(t)$, each defined by

$$s_k(t) = s(t + T_k) \quad , 0 \leq t \leq \tau_k^{\text{seg}} \tag{5.12}$$

where T_k is the position and τ_k^{seg} the duration of the k^{th} segment $s_k(t)$. Note that it is necessary that $\tau_k^{\text{seg}} \geq T_{k+1} - T_k$, with the inequality holding if the adjacent segments $s_k(t)$ and $s_{k+1}(t)$ overlap (see the final paragraph of this section).

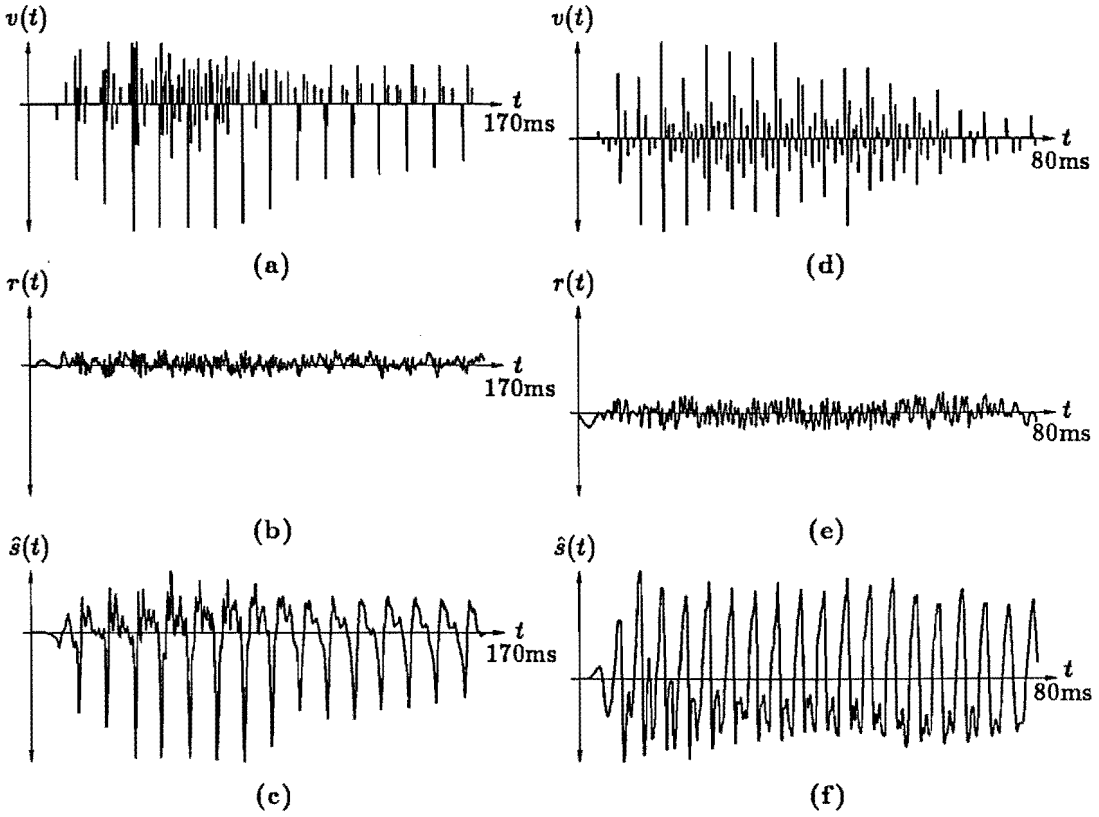


Figure 5.7. The results of applying CLEAN with a loop gain γ of 1.0, to the speech segment shown in Figs.5.3*a* and *b*. **a:** — **f:** are the same as in Fig.5.5. 175 and 450 CLEAN iterations respectively are performed for **a:** — **c:** and **d:** — **f:**, resulting in SNR = 15dB in each case. The CLEAN signals shown in *a* and *d* contain the equivalent of 940 and 2640 pulses per second respectively.

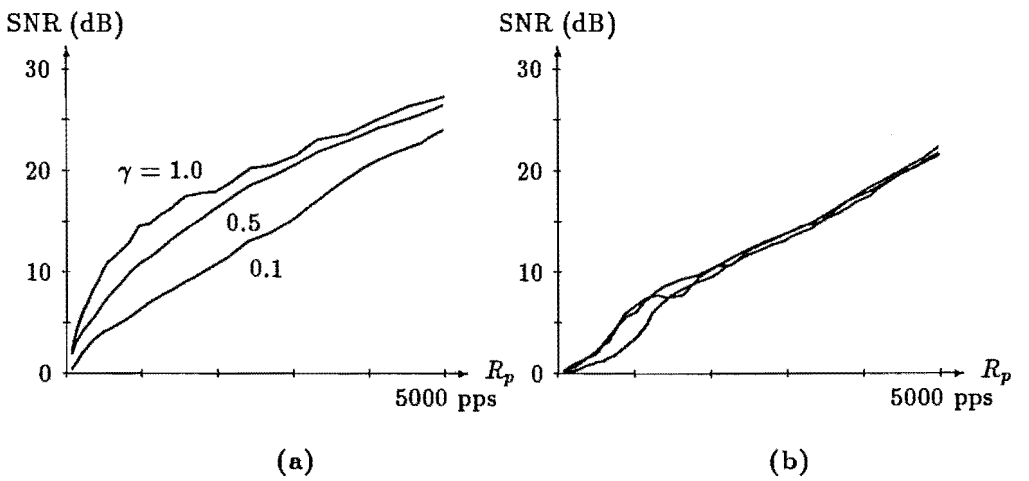


Figure 5.8. SNR as a function of the number N_p of non-zero samples in the CLEAN signal. **a:** pertains to the segment shown in Fig.5.3*a* while **b:** pertains to the segment shown in Fig.5.3*b*. Curves are shown values of the loop gain γ equal to 0.1, 0.5, and 1.0. Note that the abscissa is expressed in terms of R_p, the number of CLEAN pulses per second.

The first type of segmentation edge effect that I wish to discuss occurs because of the truncation inherent in (5.12), coupled with the finite duration of the CLEAN kernel $g(t)$. When a CLEAN pulse occurs near the end of $s_{k-1}(t)$, the shifted kernel (step 5 of the CLEAN algorithm in §5.2.3) extends beyond the end of $s_{k-1}(t)$ and into the k^{th} segment $s_k(t)$. The k^{th} segment must therefore be modified to take into account the effect of this pulse. Thus, CLEAN is performed on the modified segment $\bar{s}_k(t)$, defined by

$$\bar{s}_k(t - T_k) = s(t) - v_{k-1}(t - T_{k-1}) \odot g(t), \quad T_k \leq t \leq T_k + \tau_k^{\text{seg}}, \tag{5.13}$$

where $v_{k-1}(t)$ is the CLEAN signal for the $(k-1)^{\text{th}}$ segment.

Another edge effect arises because the CLEAN kernel $g(t)$ is non-causal (in the sense that it is non-zero for $t < 0$), which means that when a CLEAN pulse occurs near the start of a segment, the shifted kernel extends into the *previous* segment (see the above paragraph for the analogous situation for pulses near the end of a segment). In the absence of segmentation, the subtraction of these copies of the kernel may cause additional pulses to be identified by subsequent iterations of the CLEAN algorithm in that part of the signal corresponding to the previous segment. However, because of segmentation, the CLEAN signal in the $(k-1)^{\text{th}}$ segment cannot be updated when the kernel overlaps from the k^{th} segment. As a result, existing CLEAN pulses near the end of the segment $(k-1)^{\text{th}}$ may have incorrect amplitudes, and pulses that would have existed in the absence of segmentation may not be present.

Another type of segmentation edge effect occurs when the maximum magnitude sample located by step 1 of the CLEAN algorithm in §5.2.3 occurs at the end-point $s_k(\tau_k^{\text{seg}})$, but when that sample is not a local maximum of $s(t)$ (i.e. the edge of the k^{th} segment truncates a slope in $s(t)$). This results in an erroneously positioned peak, which can be avoided by testing for the possibility of such an occurrence in step 1 of the algorithm.

Because of the edge effects described above, it seems that the duration of each segment τ_k^{seg} should be as large as is convenient. However, larger segments take longer to process, in that the speech signal at the output of the CLEAN analysis and reconstruction process is delayed with respect to the input speech signal. They also require more computation, since each iteration of the CLEAN procedure must search through more samples before it locates the maximum. In order to partially allay the effects of segmentation, adjacent segments are overlapped, so that the regions in which edge effects occur (i.e. approximately half the duration of $g(t)$ from each end of the segment) are processed in both the k^{th} and $(k+1)^{\text{th}}$ segments. For the results presented in the remainder of this chapter, I employ a segment duration $\tau_k^{\text{seg}} = 25\text{ms}$ and a spacing, $\Delta T = T_{k+1} - T_k = 15\text{ms}$ between adjacent segments. This provides an overlap of 10ms.

5.2.5.4 Terminating the iterations

The model of speech invoked to explain the processing of speech by SAA and CLEAN (§4.2.1,§5.2.1) consists of a convolution of an invariant component (represented by the SAA signal) and a variant component (represented by the CLEAN signal), together with an additive "contamination" term that encompasses all aspects of the speech signal that cannot be described by the aforesaid convolution. In each iteration of the CLEAN algorithm, one of the (shifted and scaled) copies of the invariant component, from which the signal is composed, is removed from the dirty signal and an equivalent impulse is added to the CLEAN signal. After many such iterations, the dirty signal is reduced to the level of the contamination, implying that further iterations cannot improve the CLEAN signal. Since the contamination level is unknown, indirect measures must be

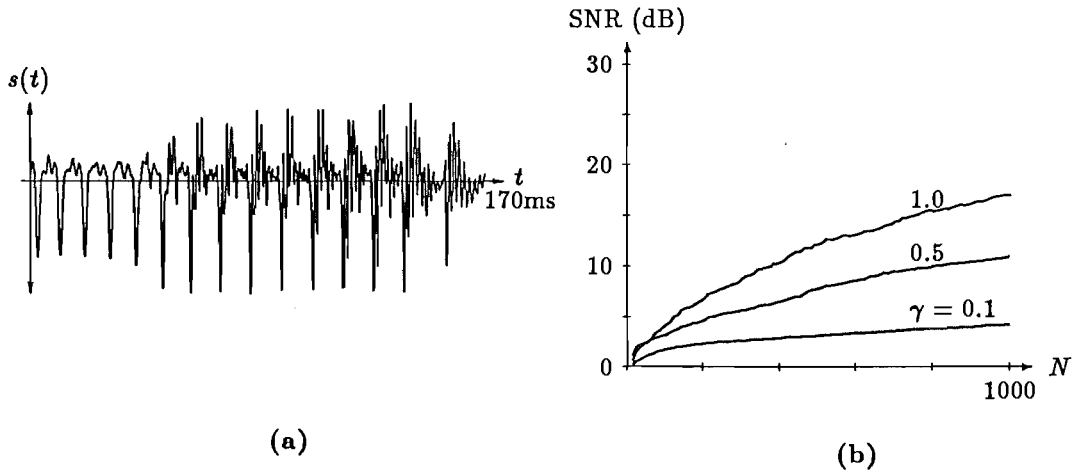


Figure 5.9. *a* A segment of speech (different from that shown in Fig.5.4*b*) from the utterance AM-RAIN1. **b:** Curves of SNR versus the number of iterations when CLEAN is applied to the segment shown in *a* with values of γ equal to 0.1, 0.5, and 1.0.

invoked to decide when to terminate the CLEAN procedure. In this section I discuss some of the methods which I have found to be useful for this purpose. It is obviously important, if one desires to minimise the number of CLEAN pulses (such as for low data rate speech encoding, §5.3), to terminate the CLEAN algorithm as soon as the dirty signal has been reduced "enough".

Perhaps the simplest approach to estimating when to terminate the CLEAN algorithm is to apply an arbitrary threshold η_{cl} on the dirty signal. The iterations are terminated when the maximum magnitude of the dirty signal (as found in step 1 of the algorithm described in §5.2.3) is less than η_{cl}, which is an estimate of the contamination level within the speech signal.

In order to adjust the threshold to account for the varying amplitude of the speech signal from syllable to syllable (§2.1.3), I set a threshold η_{cl_k} to a pre-determined proportion of the peak magnitude within each speech segment (see §5.2.5.3 for details on how the speech utterance is segmented). In addition, I set a global minimum threshold η_{cl_g} so that segments which contain inter-word silences are not unnecessarily CLEANed. The actual threshold η_{cl} for any particular segment is set to the larger of η_{cl_g} and η_{cl_k}.

One drawback with applying a threshold as described above to determine when a segment of speech has been CLEANed "enough" is that the actual contamination level may vary significantly between different segments, even when expressed as a proportion of the peak speech magnitude. As is discussed in §5.2.2, some segments of an utterance are modelled well by a convolution between $s^i(t)$ and a $s^v_m(t)$ composed of a few discrete impulses, implying that the additive contamination $s^c_m(t)$ is small for that segment. However, other segments are not so well modelled. If the contamination level is much larger than the threshold η_{cl}, the CLEAN algorithm may stagnate before it reaches the threshold. This is illustrated in Figs.5.9 which shows SNR as a function of the number of iterations of the CLEAN algorithm when it is applied to a different segment of the utterance AM-RAIN1 as that depicted in Fig.5.3*a*. The segment depicted in Fig.5.9*a* has a much higher contamination level than the one depicted in Fig.5.3*a*, as evidenced by the stagnation of the CLEAN algorithm at a lower SNR.

In order to account for the effects of stagnation, I set a limit $J_{\max}$ on the number of iterations that the CLEAN algorithm can complete for any particular segment of

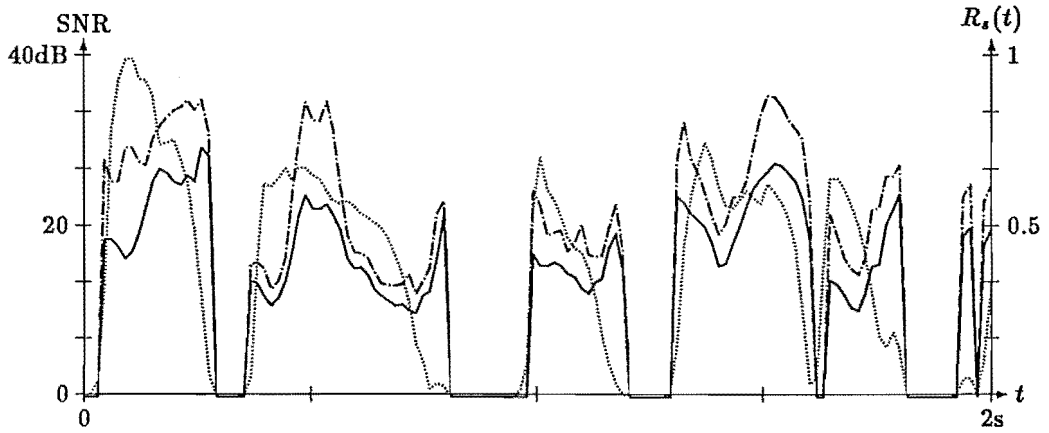


Figure 5.10. Segmental SNR during the voiced sections of the first two seconds of the utterance TF-RAIN1 (solid line). The maximum number of CLEAN pulses P_{max} in each segment was set to 30 (corresponding to an average pulse rate of 1400pps), while the maximum number of iterations J_{max} was set to 60. The thresholds η_{cl_g} and η_{cl_k} were set to zero. The dotted line represents the RMS envelope of the speech signal while the broken line represents the SNR when the pulse amplitudes are re-optimised in the manner described in §5.2.6. The average (unoptimised) SNR over the extent of the signal is 16dB.

speech. I also limit the number P_{max} of non-zero "CLEAN pulses" that the CLEAN signal (in any one segment) can be composed of. These limits act like a threshold on the dirty signal that varies according to the level of the contamination in the speech signal. This is because, for a given P_{max} or J_{max}, the level of the dirty signal is much lower for a segment with little contamination than for one with a larger contamination level. Fig.5.10 shows the variation of SNR during an utterance when these limits are applied.

Limiting the number of CLEAN pulses by means of P_{max} is attractive in the low data rate scheme described in §5.3 because it bears directly on the data rate of the encoded speech. It is thus relatively straightforward to adjust the coding strategy for different data rates.

5.2.5.5 Dealing with unvoiced speech

It is explained in in Chapter 4 (§4.2.4.5) that the voiced and unvoiced sections of a speech signal are so different in character that they must be separated and processed independently by SAA. The same reasoning applies to CLEAN processing, since it makes no sense to deconvolve a signal that represents the excitation of the voiced sections of an utterance from a section of unvoiced speech (see §5.4.3.2).

The most straightforward way to deal with the differences between voiced and unvoiced sections of speech is, first, to separate them by means of standard VUV analysis techniques (see §3.1.2), and then to perform SAA and CLEAN separately on the voiced and unvoiced sections. However, as explained in §4.2.4.5, this is often inadequate, both because of errors in the VUV classification, and because of the occurrence of sections of speech having a mixed excitation.

Another way of separating the effectively voiced and unvoiced sections of an utterance is to filter it into two frequency sub-bands. As described in §4.2.4.5 for the case of SAA processing, the high frequency sub-band contains most of the energy

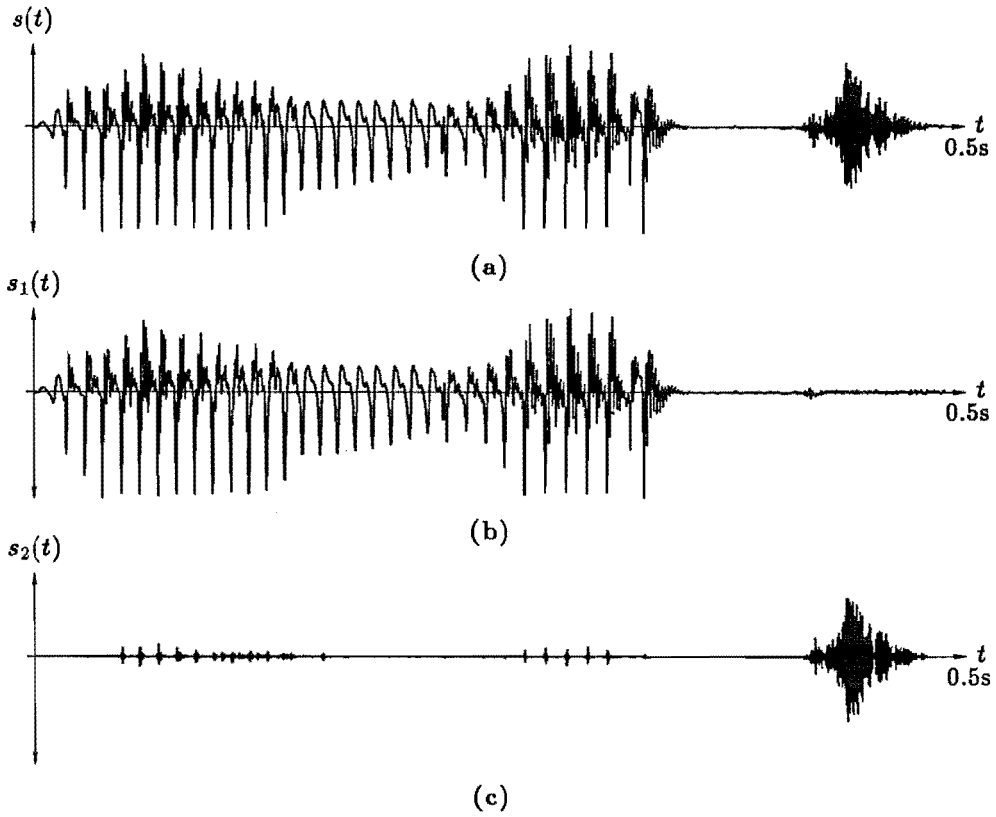


Figure 5.11. a: A section of speech (the word "raindrops") from the utterance AM-RAIN1. **b:** Low frequency sub-band 0–2.5kHz. **c:** High frequency sub-band 2.5–5kHz.

arising from the unvoiced excitation, while the low frequency band contains most of the voiced excitation. Sections of speech having a mixed excitation are represented by energy distributed between both of the bands.

Fig.5.11 *a* shows a section of speech containing both voiced and unvoiced parts. The two sub-bands are shown in Figs.5.11 *b* and *c* respectively. Performing CLEAN on these sub-bands results in the reconstructed and residual signals shown in Figs.5.12. Notice how the voiced and unvoiced sections shown here are neatly allocated to the two frequency bands, with only a small amount of energy in the low and high frequency sub-bands during the unvoiced and voiced sections of the utterance respectively. Further implications of processing speech in several sub-bands are discussed in §5.4.3.3.

5.2.5.6 Differentiation of the speech signal

The success of the CLEAN algorithm in deconvolving the SAA signal from the speech signal rests on the assumption that the peaks in the speech signal (and the dirty signal at later iterations of the algorithm) represent a "copy" of the SAA signal which is located at the same instant as the peak. As discussed in §4.2.2, these assumptions are not strictly met by actual speech signals. The glottal pulse does not contain a dominant impulsive peak, which means that highest peak in the speech signal identified in the CLEAN algorithm (§5.2.3) may not correspond to the highest peak in the glottal pulse. Pre-emphasising the speech signal by first order differentiation should therefore improve the performance of SAA (see §4.2.4.7) and CLEAN because it makes the effective excitation signal more "peaky".

Fig.5.13 shows curves of the SNR versus the pulse rate for speech that has

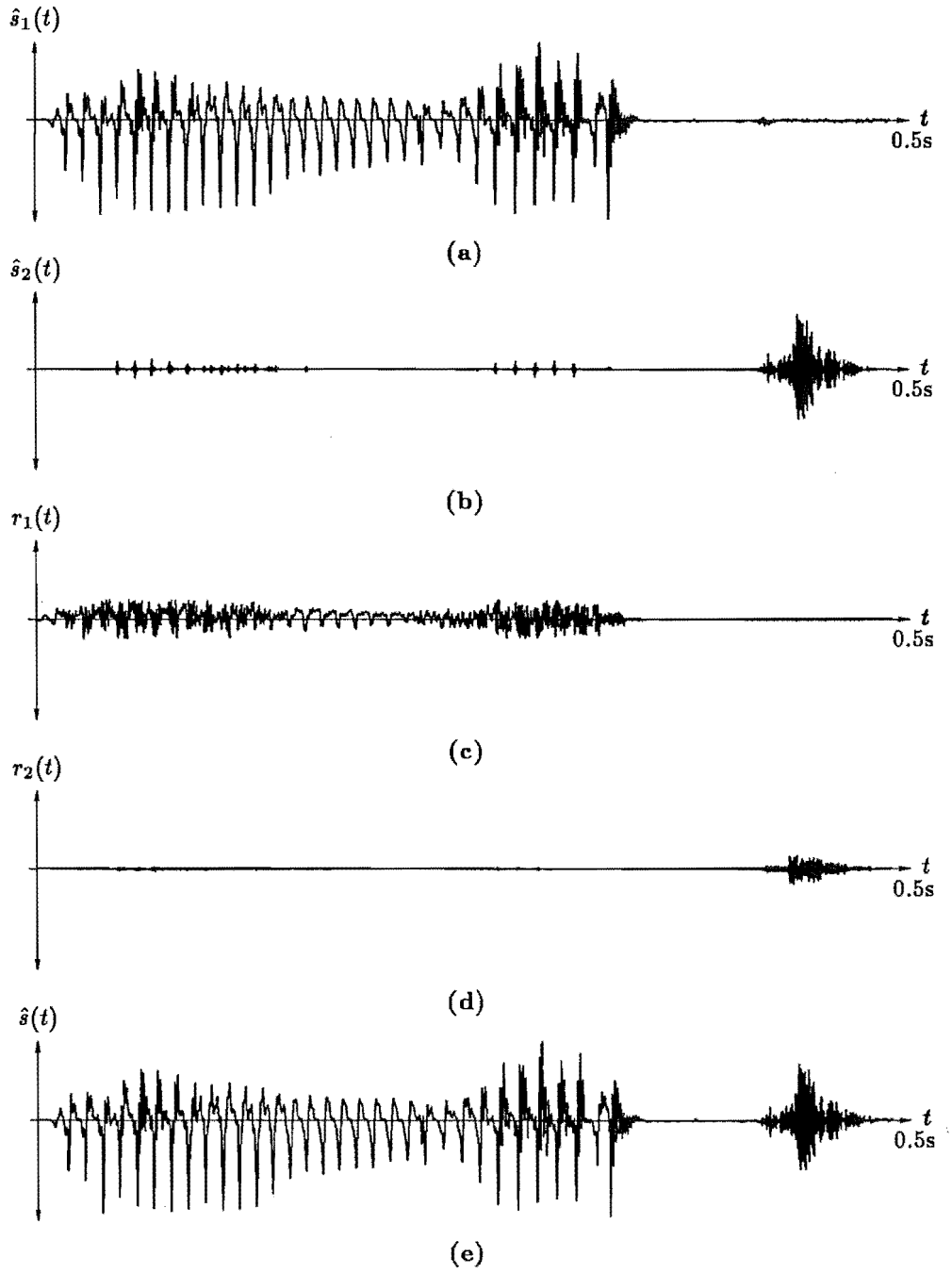


Figure 5.12. Reconstructions of the **a:** low frequency and **b:** high frequency sub-bands shown in Fig.5.11*b* and *c* respectively after they have been CLEANed with the SAA signals obtained from the respective sub-bands of the entire utterance (see Figs.4.22*a* and *b*). **c:** and **d:** Residual signals of the low and high frequency sub-bands respectively. **e:** Total reconstructed signal formed by adding together the signals shown in *a* and *b*. The average SNR of the low frequency band is 11dB, with 1400 CLEAN pulses per second. The corresponding values for the high frequency sub-band are 12dB and 1700pps respectively. $\gamma = 0.5$ in each case

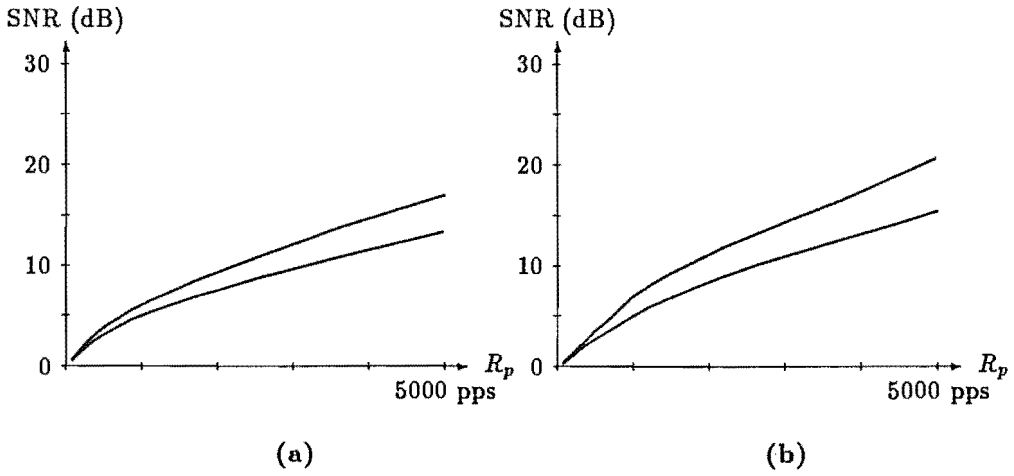


Figure 5.13. SNR versus the pulse rate for speech that has been pre-emphasised by first-order differentiation before SAA and CLEAN was performed. SAA and CLEAN was performed on the voiced sections of the utterances **a:** AM-RAIN1 and **b:** TF-RAIN1. The upper line in each graph denotes the SNR after amplitude re-optimisation (see §5.2.6). A loop gain factor $\gamma = 0.7$ was employed.

been pre-emphasised before SAA and CLEAN processing was performed. Comparing Fig.5.13 with the equivalent curves shown in Fig.5.4 indicates that pre-emphasis actually reduces the SNR at any particular pulse rate. however, the reconstructed speech sounds considerably better than without pre-emphasis. This is probably because the pre-emphasis enhances the high-frequency components so that they are better represented by the CLEAN signal (see §5.4.3 for further discussion of this point). In addition, the necessary de-emphasis of the reconstructed signal reduces the error in the high-frequency components. Because of the perceived improvement in speech quality that is obtained, I employ pre-emphasis in the low data rate speech coding scheme described in §5.3.

Another advantage of differentiating the speech signal before performing SAA and CLEAN is that the resulting SAA signal shows a tendency towards having end-points of zero amplitude. As described in §5.2.5.1, the SAA signal must be modified, so that its end-points are at zero amplitude, before it can be employed in the CLEAN algorithm. Fig.5.14 shows the SAA signals obtained from the voiced sections of two utterances, both before and after the utterances have been differentiated. In my experience, SAA signals obtained from differentiated utterances spoken by male speakers almost always have zero-amplitude end-points (such as that one shown in Fig.5.14*a*). However, SAA signals obtained from differentiated utterances spoken by female speakers usually have non-zero end-points, as is illustrated in Fig.5.14*b*. They must therefore still be modified in the manner described in §5.2.5.1.

5.2.6 Optimisation of the CLEAN pulses

In §5.2.5.2, it is noted that the form of the CLEAN signal varies markedly with changes in the loop gain factor γ. Each such signal is an approximate solution to the deconvolution problem, in that it is consistent with the given speech and SAA signals, at the specified contamination level. However, each CLEAN signal is "sub-optimal" in the sense that it is possible to adjust the pulse amplitudes to further reduce the contamination level. The reason that the simple CLEAN signal is sub-optimal is that the pulses are located sequentially. At each iteration, the amplitude of the new pulse is estimated

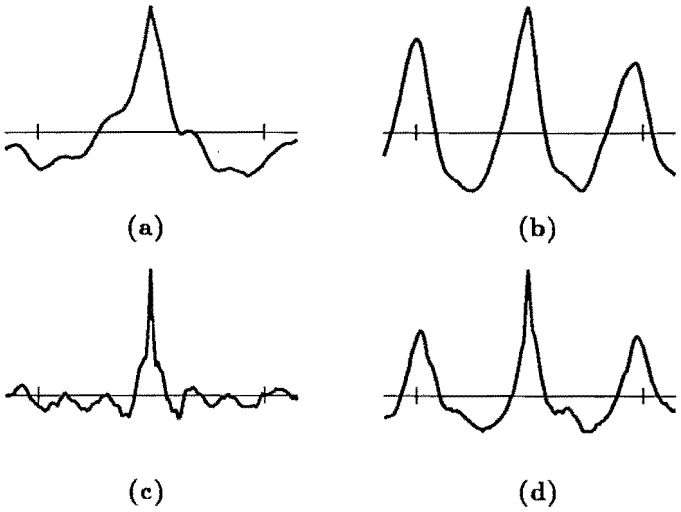


Figure 5.14. SAA signals obtained from the voiced sections of undifferentiated and differentiated utterances. Undifferentiated: a: Utterance AM-RAIN1. b: Utterance TF-RAIN1. Differentiated: c: Utterance AM-RAIN1. d: Utterance TF-RAIN1.

by considering only the speech signal and the effects on it of the previously estimated pulses. Later pulses may modify the dirty signal such as to make the current estimate of the pulse position or amplitude inaccurate. If a peak in the dirty signal reappears in later iterations at the position of a particular pulse, its amplitude is updated. However, significant improvements to r_{lev} can be obtained by *re-optimising* the pulse amplitudes after they have been located by the CLEAN algorithm.

It is convenient to present the optimisation algorithm in the sampled time domain. Readers should bear in mind, however, the implied correspondence between the sample index n and the time index t (strictly, $t = nT_s$, where $1/T_s$ is the sampling frequency).

The mean square error E between the original speech $s[n]$ and the speech signal reconstructed from the CLEAN signal is given by

$$E = \sum_{n=-\infty}^{\infty} \left[s[n] - \sum_{j=1}^{N_p} v_j g[n - p_j] \right]^2 \tag{5.14}$$

where N_p is the number of non-zero samples in the CLEAN signal, the j^{th} of which is of amplitude v_j and position p_j. Minimising E with respect to each of the v_j leads to a matrix equation which can be solved using standard techniques. For computational reasons, it is convenient to perform the amplitude optimisation on blocks of N_{opt} CLEAN pulses, holding the remaining pulses constant. Denoting the index of the first pulse in such a block as j_o, (5.14) can be re-written as

$$E = \sum_{n=-\infty}^{\infty} \left[y[n] - \sum_{j=j_o}^{j_o+N_{\text{opt}}-1} v_j g[n - p_j] \right]^2 \tag{5.15}$$

where

$$y[n] = s[n] - \sum_{i=1}^{j_o-1} v_i g[n - n_i] - \sum_{i=j_o+N_{\text{opt}}}^{N_p} v_i g[n - p_i] \tag{5.16}$$

is the signal remaining when the effects of all the CLEAN pulses that are not within the optimisation block have been removed from the speech signal.

Setting the partial derivatives of E with respect to each of the v_j to zero leads to the matrix equation

$$\sum_{k=j_o}^{N_{\text{opt}}-1} g_{p_k p_j} v_k = c_{p_j}, \quad j = j_o \ldots N_{\text{opt}} - 1 \tag{5.17}$$

where

$$g_{p_i p_j} = \sum_{n=-\infty}^{\infty} g[n - p_i] g[n - p_j] \tag{5.18}$$

is the auto-correlation of the CLEAN kernel $g[n]$ and

$$c_{p_i} = \sum_{n=-\infty}^{\infty} g[n - p_i] y[n] \tag{5.19}$$

is the cross-correlation between the CLEAN kernel $g[n]$ and the modified speech signal $y[n]$ for the current optimisation block. Standard matrix solving techniques can be invoked to solve (5.17) and thereby obtain "optimised" values for the CLEAN pulses v_k.

Because $g[n]$ is non-zero for $n_g^- < n < n_g^+$ only, the computational effort required to evaluate (5.16) can be reduced since only those CLEAN pulses whose positions p_i satisfy $p_{j_o} - n_g \leq p_i < p_{j_o}$ or $p_{(j_o+N_{\text{opt}})} \leq p_i \leq p_{(j_o+N_{\text{opt}}-1)} + n_g$, where $n_g = n_g^+ - n_g^-$, need be subtracted. In practice, it is easier to form $y[n]$ by *adding* the effects of the pulses within the optimisation block to the residual $r[n]$, since the residual is available from the CLEAN algorithm. A practical algorithm, which performs the CLEAN pulse amplitude re-optimisation, proceeds according to the following sequence of steps, with the initial residual $r^{(0)}[n]$ set equal to the residual remaining after the CLEAN algorithm (§5.2.3) has been applied to all segments (see §5.2.5.3) of the utterance being processed.

1. Compute the autocorrelation $\text{gg}[i] = \sum_{n=-\infty}^{\infty} g[n] g[n - i], \quad i = 0 \ldots n_g.$

2. For each block of N_{opt} CLEAN pulses, perform the following five operations.

   (a) Compute the modified speech signal for the k^{th} block:

$$y^{(k)}[n] = r^{(k-1)}[n] + \sum_{j=j_{ok}}^{j_{ok}+N_{\text{opt}}-1} v_j g[n - p_j], \quad p_{j_{ok}} - n_g < n < p_{(j_{ok}+N_{\text{opt}}-1)} + n_g. \tag{5.20}$$

   (b) Compute the cross-correlation vector $c_{p_j}, \quad j = j_{ok} \ldots j_{ok} + N_{\text{opt}} - 1$, according to (5.19).

   (c) Isolate the appropriate elements of gg to form $g_{p_i p_j}, \quad i, j = j_{ok} \ldots j_{ok} + N_{\text{opt}} - 1$:

$$g_{p_i p_j} = \text{gg}[p_i - p_j]. \tag{5.21}$$

   (d) Obtain the updated values v_j' of the CLEAN pulse amplitudes by solving (5.17) with $v_j' = v_j$.

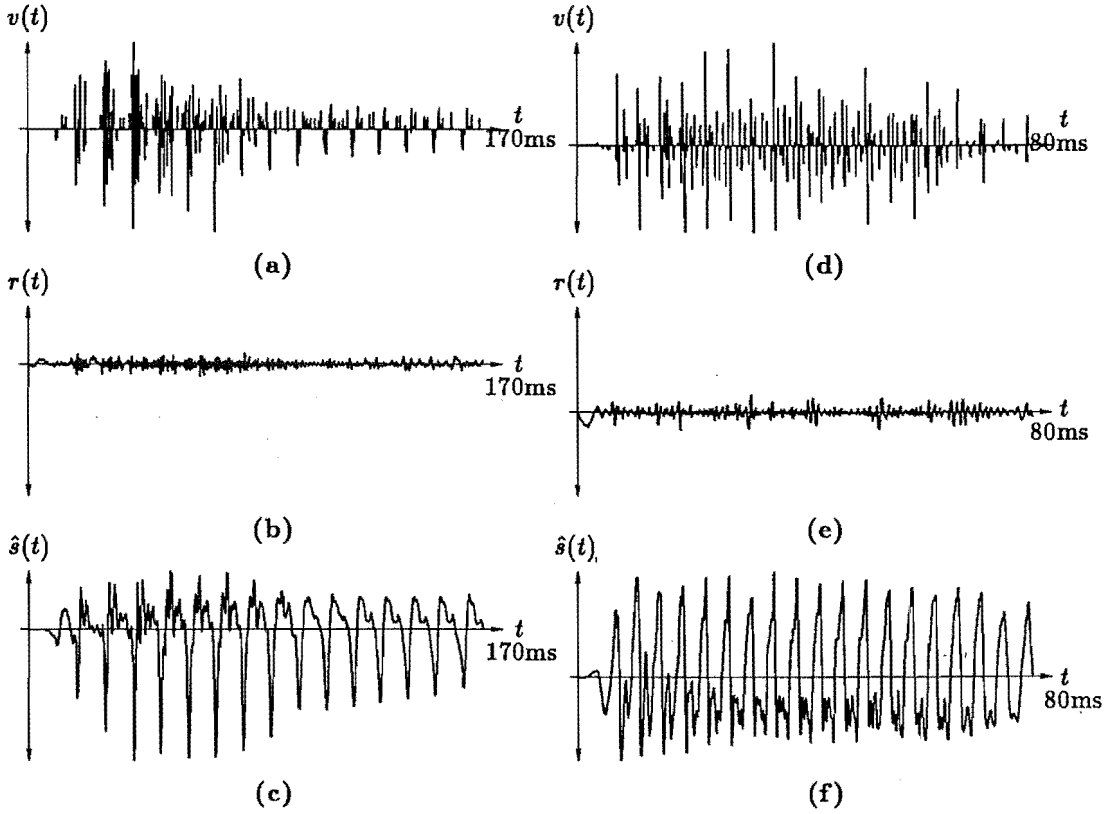


Figure 5.15. Re-optimising the CLEAN pulse amplitudes. **a:**, **d:** CLEAN signals corresponding to those shown in Fig.5.6*a* and *d* respectively but after the pulse amplitudes have been optimised. **b:**,**e:** Residual, and **c:**,**f:**, reconstructed speech signals. The SNR after optimisation is 21dB for each segment, compared with 15dB for the unoptimised equivalents.

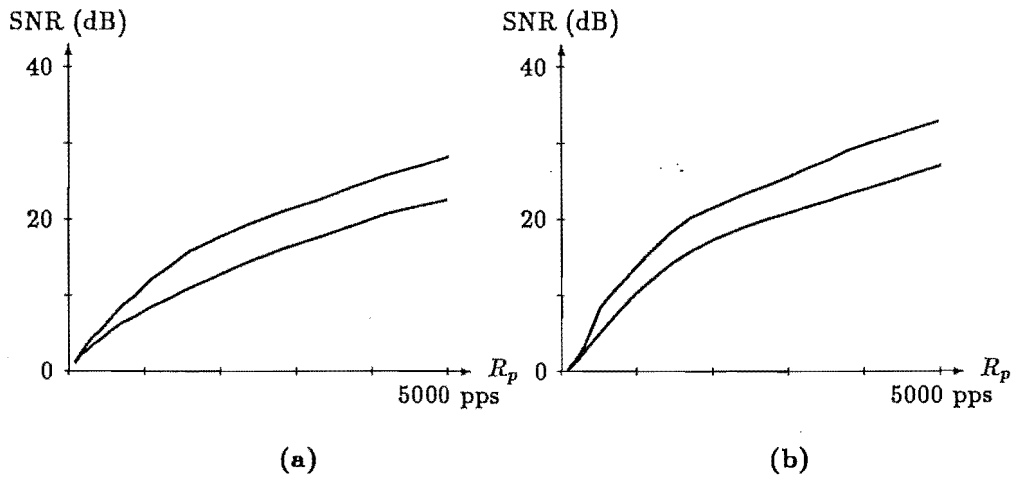


Figure 5.16. Comparison between the SNR of optimised and unoptimised CLEAN. The upper curve in each figure represents the SNR after optimisation while the lower curve represents that before optimisation. Utterances **a:** AM-RAIN1 and **b:** TF-RAIN1. A loop gain factor $\gamma = 0.7$ was employed. Note that these figures are drawn with a vertical scale of 0–40dB instead of 0–30dB as on the other SNR graphs.

(e) Update the residual signal to take into account the optimised CLEAN pulses:

$$r^{(k)}[n] = r^{(k-1)}[n] - \sum_{j=j_{ok}}^{j_{ok}+N_{\rm opt}-1} v'_j g[n-p_j], \qquad p_{j_{ok}} - n_g < n < p_{(j_{ok}+N_{\rm opt}-1)} + n_g. \tag{5.22}$$

Figs.5.15*a* and *d* shows the CLEAN signals obtained after optimising the CLEAN pulses shown in Figs.5.6*a* and *d* respectively. The resulting reconstructed speech signals are shown in Figs.5.15*c* and *f*. The improvement in SNR over the unoptimised versions are indicated by the curves shown in Fig.5.16. These graphs, together with the curves of the variation of SNR during an utterance shown in Fig.5.10, indicate that pulse amplitude re-optimisation provides roughly 6dB improvement to the SNR of the reconstructed speech.

The number of pulses $N_{\rm opt}$ that can be optimised in each block is limited by the difficulties inherent in solving large matrix equations. In addition, it does not make sense to simultaneously optimise pulses that are widely separated, since they are practically independent. Computational experience suggests that $N_{\rm opt}$ can be as large as 100 without incurring any instabilities in the matrix solution. For the results presented in this chapter, $N_{\rm opt}$ is set equal to the smaller of 100 or $4P_{\rm max}$ (where $P_{\rm max}$ is the number of pulses found in each CLEAN segment of speech, see §5.2.5.4). This means that the pulses are optimised over an interval of approximately 4–5 times the extent n_g of $g[n]$ (since n_g is slightly smaller than the spacing between adjacent segments).

5.3 Low data rate speech encoding via SAA and CLEAN

As indicated by the results presented in §5.2, the CLEAN signal obtained from a speech utterance consists of non-zero "CLEAN pulses" interspersed with zero-valued samples. Depending on the utterance and the level to which the dirty signal is reduced by CLEAN, the number of CLEAN pulses ranges from 500–3000 per second. This is considerably less than the total number of samples in the speech signal (10000 per second for the utterances employed here) implying that the CLEAN pulses can be used as a low data rate representation of the speech signal. However, because the CLEAN pulses are not uniformly spaced, both their amplitudes and positions need to be encoded. In order to take full advantage of the benefits implied by the reduced number of pulses, efficient means of encoding the pulse positions and amplitudes must be employed.

§5.3.1 describes the methods by which the CLEAN pulses are encoded, while §5.3.2 gives details of how the SAA and CLEAN analysis techniques described elsewhere in this and the previous chapter are employed for low data rate speech encoding. §5.3.3 presents results pertaining to the performance of the scheme.

5.3.1 Encoding the CLEAN signal

A CLEAN signal is completely specified by the amplitudes and positions of all of its constituent CLEAN pulses. For encoding purposes, it is convenient to separately consider the information pertaining to the amplitudes of the pulses, and that which defines their positions. Following this rationale, I here represent a CLEAN signal by a set of amplitudes $\{v_j,\ j = 1,\ldots,N_{\rm tot}\}$ and a set of positions $\{p_j,\ j = 1,\ldots,N_{\rm tot}\}$, where the $j^{\rm th}$ (out of $N_{\rm tot}$ in total) CLEAN pulse has amplitude v_j and is located at time p_j. §5.3.1.1 describes how the pulse amplitudes are encoded, while §5.3.1.2 is concerned with the encoding of the pulse positions.

5.3.1.1 Encoding amplitudes

The elements of the set of pulse amplitudes $\{v_j\}$ are real numbers, which means that they must be quantised before they can be digitally encoded. Efficient quantisation and encoding requires that the statistical structure of the quantities to be encoded be reflected in the coding scheme (cf. Hamming, 1980). Two useful descriptors of the statistical nature of a list of numbers are its probability distribution function (pdf) and autocorrelation function. The pdf (§1.3.3) is approximated by a histogram showing the relative frequency of occurrence of each (quantised) amplitude value in the set $\{v_j\}$. The normalised autocorrelation function $c_a[k]$ of the ordered set of numbers $\{a_j\}$ is defined by

$$c_a[k] = \frac{\langle \hat{a}_j \hat{a}_{j-k} \rangle_j}{\langle \hat{a}_j^2 \rangle_j} \tag{5.23}$$

where $\hat{a}_j = a_j - \bar{a}$ and $\bar{a}$ is the ensemble mean of $\{a_j\}$. The autocorrelation gives an average indication of how much any particular pulse amplitude is related to the amplitudes of nearby pulses.

Fig.5.17 shows histograms of the pulse amplitudes for various CLEAN signals. These show that the CLEAN pulses are predominantly of low amplitude. The autocorrelation functions $c_v[k]$ of the sets of pulse amplitudes from several CLEAN signals are shown in Fig.5.18. These indicate that the amplitudes of nearby pulses in the CLEAN signal are somewhat correlated, but that this correlation falls off rapidly with increasing k.

Because the statistics of the pulse amplitudes as revealed by the histograms and correlation functions shown in Figs.5.17 and 5.18 are somewhat similar to those describing speech signals (cf. Rabiner and Schafer, 1978, §5.2), the techniques used to quantise speech signals (see §3.5.1) can be usefully invoked to quantise the pulse amplitudes. I employ a simple adaptive quantisation scheme where the quantisation levels in each block of N_b pulse amplitudes are adjusted to the maximum magnitude pulse within that block. This is the same scheme described by Kroon and Deprettere (1988) to quantise the amplitudes of their multi-pulse excitation sequences. The maximum magnitude pulse within each block, called the quantisation *gain*, is denoted by G_k. Each amplitude value within the block is normalised, by dividing it by G_k, and then uniformly quantised with q_a bits. The gain is also uniformly quantised, with q_g bits, and transmitted along with each block. The average number of bits q_{av} required to code each pulse amplitude is therefore

$$q_{av} = q_a + q_g/N_b. \tag{5.24}$$

The quantisation parameters q_a, q_g, and N_b affect both the resulting data rate of the encoded speech and the quality of the reconstructed speech. Hence the particular values that are given to these parameters must be determined according to the required data rate and quality. §5.3.3 describes the results obtained when various values are employed for q_a and q_g. Computational experience suggests that $N_b = 6$ is the best value for this parameter. Larger values for N_b result in each block containing pulses whose amplitudes are significantly less correlated (as suggested by the low correlation values shown in Fig.5.18 for $k > 6$), while smaller values increase q_{av}.

5.3.1.2 Encoding the pulse positions

The sequence of CLEAN pulse positions $\{p_j\}$ constitutes a point process (cf. Cox and Isham, 1980). It is convenient to represent this sequence by the set of inter-pulse intervals $\{I_j\}$, where

$$I_j = p_j - p_{j-1}, \quad j = 1 \ldots N_{\text{tot}} \tag{5.25}$$

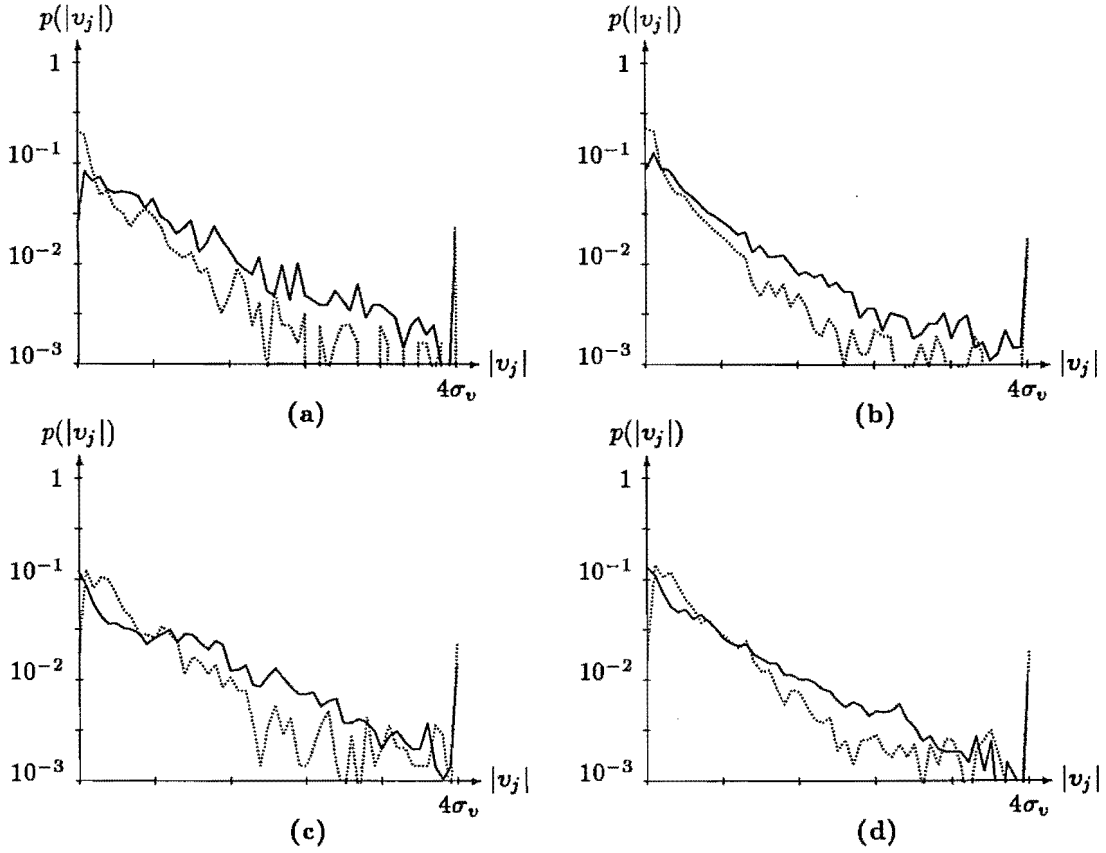


Figure 5.17. Histograms of the distribution of pulse amplitudes in various CLEAN signals. The histograms are plotted on a logarithmic scale, with the solid line pertaining to the low frequency sub-band and the dotted line the high frequency sub-band. a: AM-RAIN1, 420 and 250pps in the low and high frequency sub-bands respectively. b: AM-RAIN1, 1100 and 640pps. c: TF-RAIN1, 520 and 250pps. d: TF-RAIN1, 1300 and 620pps. Note that the large value of $p(|v_j|)$ on the extreme right of each graph actually represents $P(|v_j| > 4\sigma_v)$, where σ_v is the standard deviation of $\{v_j\}$.

and $p_0 = 0$. Note that the intervals I_j are positive integers, in units of the sampling interval of the CLEAN signal. An alternative representation of the sequence of pulse positions is as a binary sequence, with 1's representing the pulse positions (the non-zero samples in the CLEAN signal) and 0's representing the intervening zero-valued samples. Because the number of CLEAN pulses is relatively small (i.e. $P(1) << P(0)$), this binary sequence consists of "runs" of 0's terminated by 1's. The interval values I_j defined in (5.25) are equal to the length of the runs plus one.

A point process is described by the conditional probability

$$P(I_k|I_{k-1}, I_{k-2}, \ldots) \tag{5.26}$$

of a particular interval I_k occurring after a history of intervals $\{I_{k-1}, I_{k-2}, \ldots\}$ (Cox and Isham, 1980). If each interval is independent of the previous history, (5.26) reduces to $p(I_k)$, the zeroth-order pdf of $\{I_j\}$ (Hamming, 1980, §5.2). The autocorrelation $c_I[k]$ of the sequence of intervals $\{I_j\}$ gives an indication of how much the past terms in (5.26) need to be taken into account to adequately represent the sequence. Fig.5.19 shows the first few autocorrelation lags of the sets of intervals obtained from the utterances AM-RAIN1 and TF-RAIN1. For most lags $k > 1$, $|c_I[k]| < 0.05$, which suggests that $\{I_j\}$ is effectively independent of its past history and thus is adequately characterised by its pdf $p(I_j)$. Note that the small regularly spaced peaks in $c_I[k]$ result from the pitch

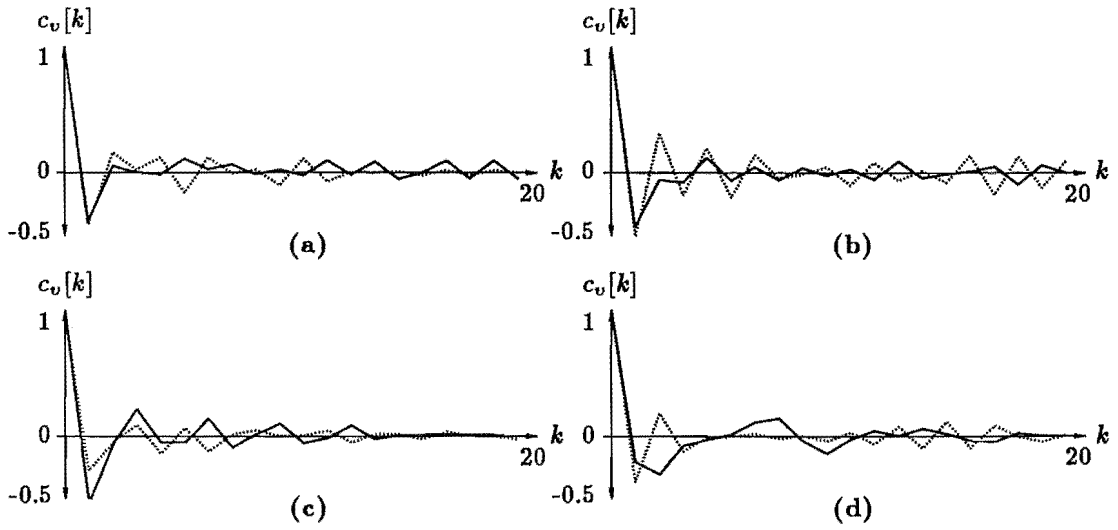


Figure 5.18. First 20 lags of the autocorrelation formed from ensembles of CLEAN pulse amplitudes. The solid line represents the low frequency sub-band and the dotted line the high-frequency sub-band. **a:** AM-RAIN1, 420 and 250pps **b:** AM-RAIN1, 1100 and 640pps. **c:** TF-RAIN1, 520 and 250pps. **d:** TF-RAIN1, 1300 and 620pps.

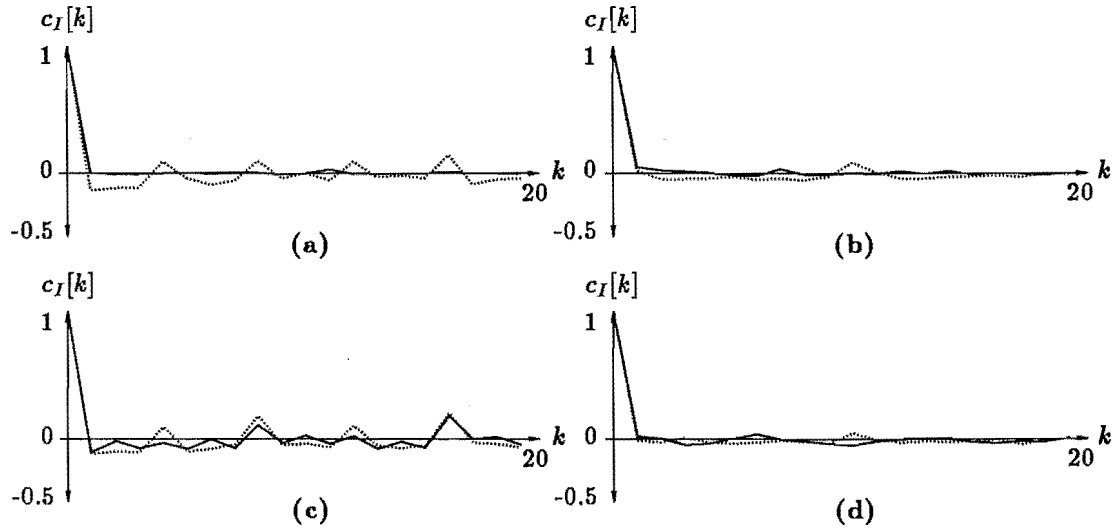


Figure 5.19. First 20 lags of the autocorrelation formed from ensembles of CLEAN pulse intervals. The solid line represents the low frequency sub-band and the dotted line the high-frequency sub-band. **a:** AM-RAIN1, 420 and 250pps **b:** AM-RAIN1, 1100 and 640pps. **c:** TF-RAIN1, 520 and 250pps. **d:** TF-RAIN1, 1300 and 620pps.

periodicity of the speech signals. A more sophisticated model of the interval process should therefore also contain a term representing this periodicity.

Histograms of the interval distributions are shown in Fig.5.20 for sets of intervals taken from several different CLEAN signals, with pulse rates ranging from 250–1200 pps. These histograms are plotted on a logarithmic scale to emphasise the almost exponential decrease in the probability of occurrence as the interval increases.

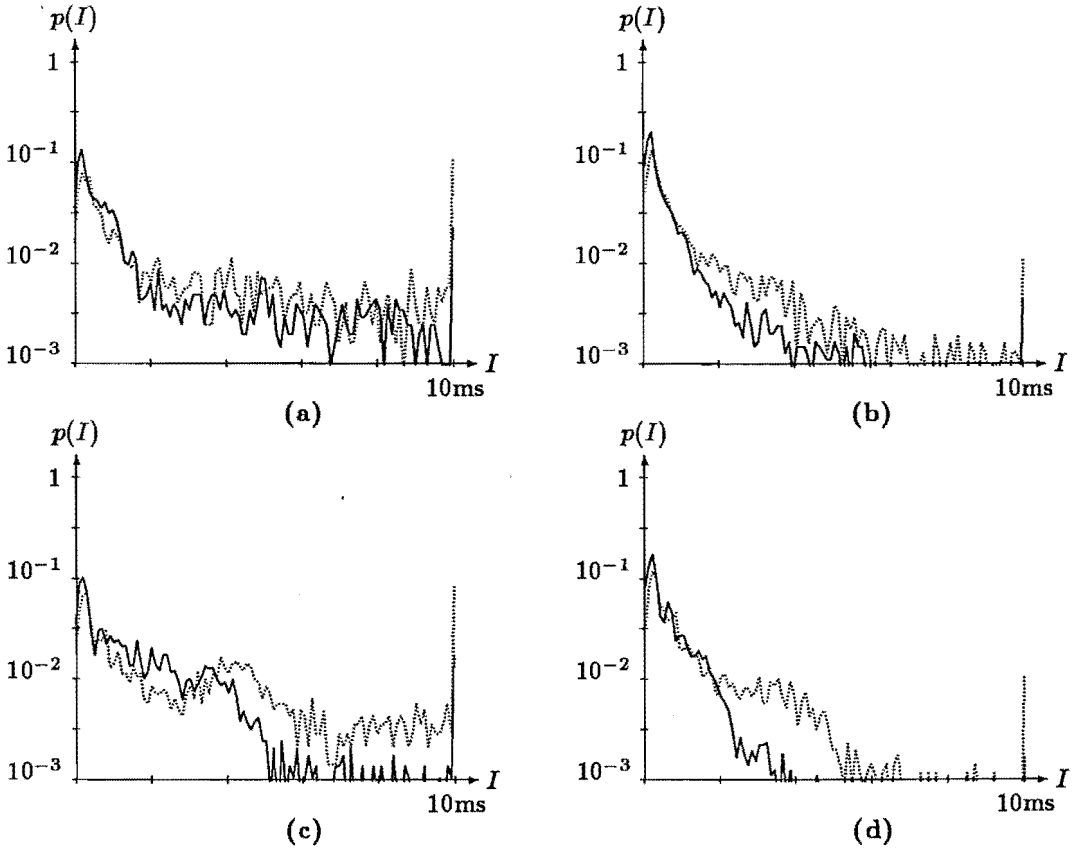


Figure 5.20. Histograms of the distribution of pulse intervals in several CLEAN signals. The solid line represents the low frequency sub-band and the dotted line the high-frequency sub-band. **a:** AM-RAIN1, 420 and 250pps **b:** AM-RAIN1, 1100 and 640pps. **c:** TF-RAIN1, 520 and 250pps. **d:** TF-RAIN1, 1300 and 620pps. Note that the large value of $p(I_j)$ on the extreme right of each graph actually represents $P(I_j > 10ms)$.

Because the probability of shorter intervals occurring is so much greater than that for longer intervals (refer to the histograms shown in Fig.5.20), efficient coding philosophy demands that they be encoded with fewer bits (cf. Shannon, 1948). A particularly simple method of encoding the intervals is to employ run-length coding (Hamming, 1980, §5.9). Run-length coding is so named because of the correspondence noted in the first paragraph of this section between the interval sequence $\{I_j\}$ and the "runs" of 0's occurring in a binary sequence when $P(1) << P(0)$.

Run-length coding is implemented by setting a fixed block-length b_q which is the number of bits that comprise the basic code word. Such a code word can directly represent intervals between 1 and $2^{b_q} - 1$ samples by means of standard binary interpretations of the bit patterns. The 2^{b_q}th bit pattern represents a "carry-on" code, indicating that another code word follows. This scheme implies that intervals between 1 and $2^{b_q} - 1$ samples require b_q bits to code, those between 2^{b_q} and $2(2^{b_q} - 1)$ require $2b_q$ bits, and so on for longer intervals. The block length b_q required to code any particular set of intervals obviously depends on the pdf of that set, with longer block lengths being required to efficiently code the intervals at lower pulse rates (when the average interval is greater — see Fig.5.20). For the results presented in §5.3.3, I indicate the block-length that is employed for each pulse rate.

The "compression ratio" CR of a run-length code, assuming independence be-

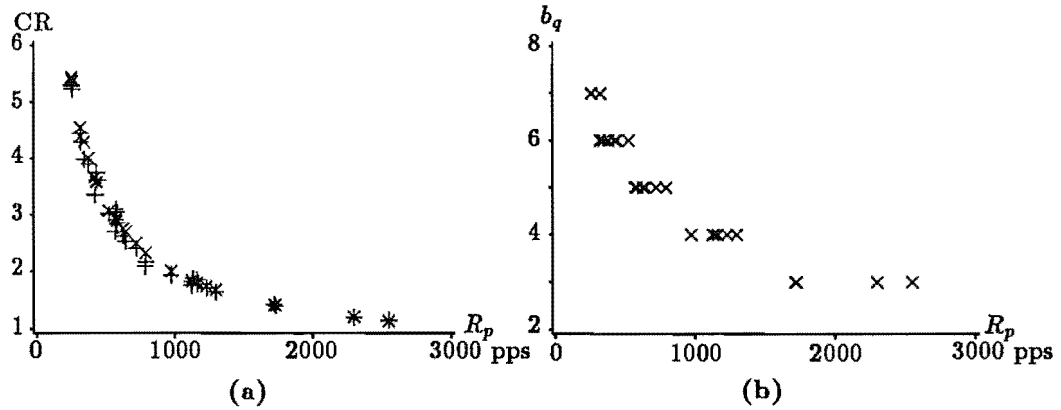


Figure 5.21. Compression ratios achieved by run-length coding of the CLEAN pulse intervals. **a:** Theoretical (×) and experimental (+) values of CR plotted against the pulse rate. **b:** Optimum block size (that which maximises the experimental value of CR) versus pulse rate. The experimental values of CR are computed from CLEAN signals obtained from utterances AM-RAIN1 and TF-RAIN1.

tween each interval, is given by (Hamming, 1980)

$$\mathrm{CR} = \frac{1 - p^{2^{b_q}-1}}{b_q(1-p)} \tag{5.27}$$

where $p = P(0)$, the probability of a 0 in the binary sequence referred to in the first paragraph of this section. Values of CR for various pulse rates are graphed in Fig.5.21*a*, together with experimentally obtained values for several utterances. The block size b_i that was found to offer the largest value of CR at each pulse rate is graphed in Fig.5.21*b*. The experimentally obtained values of CR are close to those obtained via (5.27), indicating that the assumptions of independence between members of $\{I_j\}$ are met reasonably well by the sequences of CLEAN intervals.

Note that run-length coding becomes inefficient at very high pulse rates. This is because the average interval becomes comparable in size to the block length. When the block length $b_q = 1$, the encoded signal is identical to the binary sequence referred in the first paragraph of this section. This implies that the maximum data rate (in bit/s) required to encode the pulse positions is equal to the sampling frequency (in Hz) of the CLEAN signal (10kHz here). The maximum pulse rate occurring in the results presented in Fig.5.21 is 2500pps, which, when coded with a block length $b_q = 3$, requires a data rate of 8.6 kbits/s.

While run-length coding is not as optimal as other methods such as Huffman coding (Hamming, 1980), the codes that result are generated in multiples of a fixed-length block. They can therefore be easily interlaced with the codes representing the pulse amplitudes (§5.3.1.1). In addition, the computational and memory buffering requirements are less than for "optimal" variable-length coders (cf. Mark and Todd, 1981). Further details of run-length coding are given by Hamming (1980, §5.9) and Jayant and Noll (1984, Chapter 10).

5.3.2 A practical speech encoding scheme

Fig.5.22 shows the block diagram of the complete SAA/CLEAN low data rate speech encoding scheme. This was implemented on the departmental VAX computer system (see §1.3.4). Refering to Fig.5.22, the speech signal is first separated into two sub-bands.

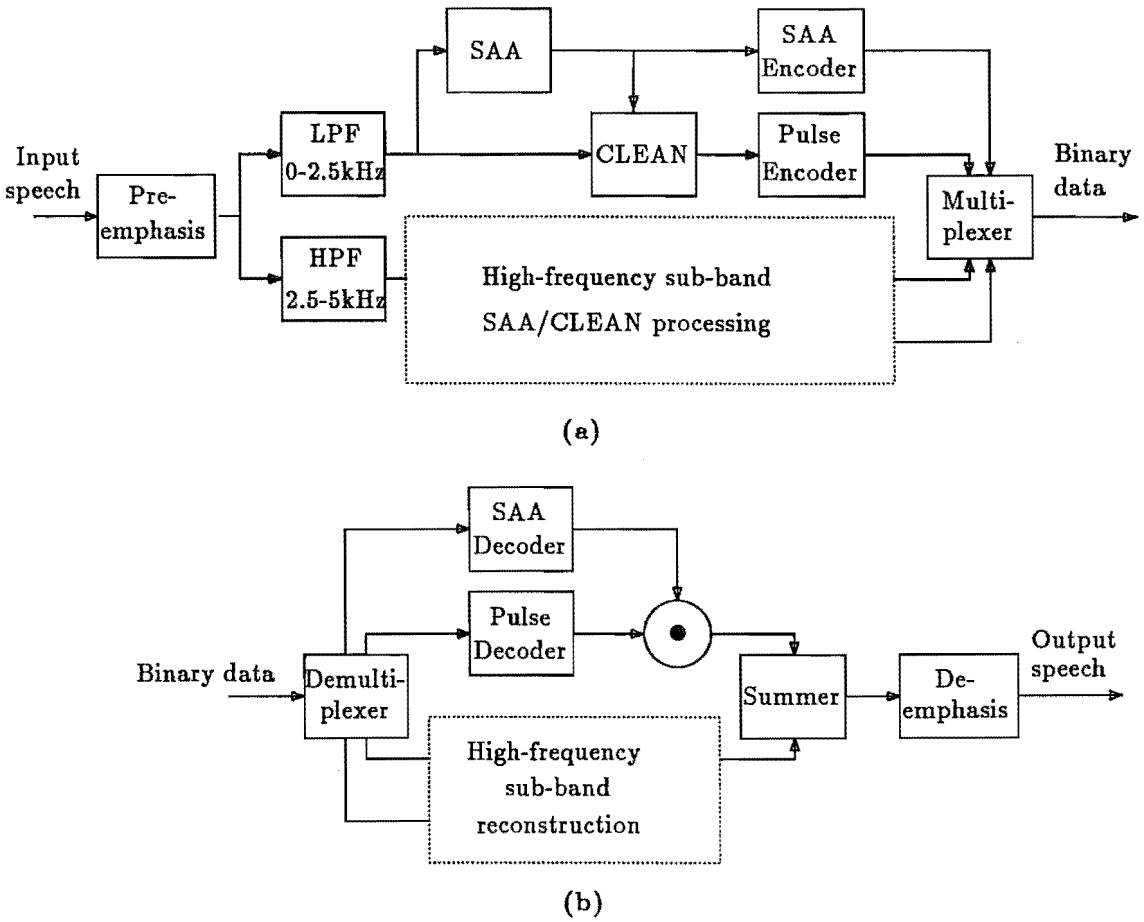


Figure 5.22. Block diagram of the complete SAA/CLEAN low data rate speech encoding and reconstruction scheme. **a:** Speech encoder. **b:** Speech decoder.

All the following processing is then performed on each sub-band. The SAA processing proceeds according to the algorithm detailed in §4.2.3 and §4.2.4 of Chapter 4. The SAA signals are modified, as described in §5.2.5.1, so that their end-points are of zero amplitude. CLEAN is then performed in the manner described in §5.2.3 through §5.2.6. The resulting CLEAN pulses are encoded in the way described in §5.3.1.1 and §5.3.1.2. Table 5.2 shows the number of bits employed for each sub-band at various data rates. Note that the higher band is encoded at a lower date rate, both by restricting the number of pulses to a smaller value than the low frequency band, and by encoding the pulse amplitudes with fewer bits.

The SAA signals, which are computed once for each sub-band of an utterance, are stored separately from the CLEAN pulses. The SAA signals require only some 800 bits to code (if the SAA signal is of duration 100 samples and is quantised to 8 bits) which, since it is averaged over the duration of the utterance, adds little to the total data rate.

The reconstructed speech utterance is formed by, firstly, individually reconstructing the two sub-bands in the manner described in §5.2.4 and, secondly, adding them together.

Name	R_p (pps)		q_a (bits)	R (bit/s)
	Low	High	Low	
TC7KB	522	253	2	7133.5
AC7KB	424	256	2	6587
TC11KB	974	378	2	11082
AC11KB	789	345	2	10102
TC13KB	974	378	4	13030
AC12KB	789	345	4	11530
TC17KB	1297	626	4	17300.5
AC16KB	1124	644	4	16296

Table 5.2. Coding details for SAA/CLEAN at various data rates. The value of q_a for the high frequency sub-band is 2 for all utterances. Other test utterances were also employed in the quality evaluations. These were formed by varying q_a between 2 and 4 (see Fig.5.29).

Name	R_p (pps)	q_a (bits)	FR_{LPC}	R (bit/s)
TM8KB	571	2	100	8242.5
AM8KB	570	2	100	8455
TM10KB	571	4	100	9384.5
AM10KB	570	4	100	9595
TM14KB	1135	4	100	13987.5
AM14KB	1129	4	100	14160.5
TM18KB	1160	4	200	17792
AM18KB	1163	4	200	18033.5

Table 5.3. Coding details and subsequent data rate for each of the utterances processed by MP-LPC that are compared with the SAA/CLEAN encoded utterances. FR_{LPC} is the frame-rate at which the LPC filter coefficients are updated. 10 LPC coefficients are computed at each frame, each set being encoded with 36 bits (Kroon and Deprettere, 1988).

5.3.3 Performance evaluation

In order to evaluate the performance of the SAA/CLEAN speech encoding technique described in this chapter, I have employed several of the speech quality measures introduced in §3.5.3. I applied these to speech signals encoded at various data rates and also to speech processed by the multi-pulse LPC (MP-LPC) technique (see §3.5.2.4). The latter is a well-known speech encoding scheme that has many similarities to the SAA/CLEAN technique (see §5.4.2), making comparison of the two schemes worthwhile.

Results are presented in §5.3.3.1 through §5.3.3.3 of the processing of two utterances, from a male and a female speaker. Table 5.2 shows details of the pulse rate, quantisation of pulse amplitudes, and consequent data rate for the utterances. The same two utterances were also processed by the "optimal" MP-LPC technique of Singhal and Atal (1989). Table 5.3 gives the coding details and data rates for these utterances. §5.3.3.1 and §5.3.3.2 respectively present the results of "objective" and "subjective" evaluations of the speech quality. §5.3.3.3 then discusses some of the implications of these results, and presents example spectrograms of the re-synthesised speech signals.

5.3.3.1 Objective evaluation of encoder performance

In order to objectively evaluate the quality of the reconstructed utterances processed by the SAA/CLEAN technique described in this chapter, several of the measures of speech quality introduced in §3.5.3.2 are invoked. Results are presented here for both the SAA/CLEAN and MP-LPC techniques, with test utterances encoded at various data rates as detailed in Tables 5.2 and 5.3 respectively. Note that the various quality measures employed here are described in detail in §3.5.3.2.

The SNR indicates the accuracy with which the waveform of the reconstructed signal matches the original. Because waveform distortions are not always perceptually relevant (§2.2.2.2), the SNR is not a reliable indicator of speech quality (§3.5.3.2). However, since both the SAA/CLEAN and MP-LPC techniques attempt to minimise the waveform error, it seems relevant to present the SNR of the reconstructed utterances. However, it is important to remember that both MP-LPC and SAA/CLEAN introduce certain perceptually insignificant waveform distortions. MP-LPC minimises a spectrally weighted error criterion which means that the perceptual effect of the error is less than the SNR would suggest. The waveform in SAA/CLEAN is distorted somewhat because of the sub-band filters. In addition, both techniques involve an initial pre-emphasis of the speech signal, and since the subsequent de-emphasis is implemented as a "leaky integrator", further waveform distortions result.

Fig.5.23 shows the segmental SNR during one of the processed utterances. Note the much greater variation in SNR in the SAA/CLEAN encoded utterance (Fig.5.23 *a*) than in the utterance that has been processed by MP-LPC. This variation arises because of the varying degrees to which different segments are modelled by the SAA/CLEAN model (see §5.2.5.4 and §5.4.2). The SNR of the MP-LPC encoded utterance is at a more constant level because the LPC filter is updated for each segment. Comparing the SNR curve shown in Fig.5.23*a* with the equivalent (but unquantised) curve shown in Fig.5.10 indicates that the various ancillary processing steps involved in the practical speech encoding scheme (i.e. separation into sub-bands, pre-emphasis, quantisation, and de-emphasis) significantly reduces the SNR of the reconstructed speech. However, as noted above, not all of this added distortion is perceptually relevant.

Figs.5.24*a* and *b* show the average segmental SNR for utterances processed by SAA/CLEAN and MP-LPC respectively as a function of data rate. The segmental SNR is computed as described in §3.5.3.2, with segment lengths of 20ms.

Other objective quality measures are based on the short term spectral or cepstral (§3.3) error in the reconstructed signal. Here I present results using log spectral and cepstral distortion measures (§3.5.3.2). Fig.5.25 shows the average log spectral distortion (see (3.53) in §3.5.3.2) as a function of data rate for utterances processed by SAA/CLEAN and MP-LPC respectively. Curves displaying the variation with data rate of the cepstral distance measure of Kitawaki *et al.* (1988) (see (3.52) in §3.5.3.2) are shown in Fig.5.26. The results of the SNR, log spectral, and cepstral distance measures shown here all indicate that MP-LPC is better than SAA/CLEAN at the lower data rates(<12kbit/s), but that the two methods give comparable results at the higher rates.

5.3.3.2 Subjective quality evaluation

Informal listening tests indicated that the MP-LPC speech has a somewhat "clearer" quality than SAA/CLEAN encoded speech, but suffers from the presence of occasional "clicks". In order to obtain a quantitative measure of the subjective quality of the re-synthesised speech, a category judgement test (see §3.5.3.1) was performed on the SAA/CLEAN and MP-LPC encoded utterances.

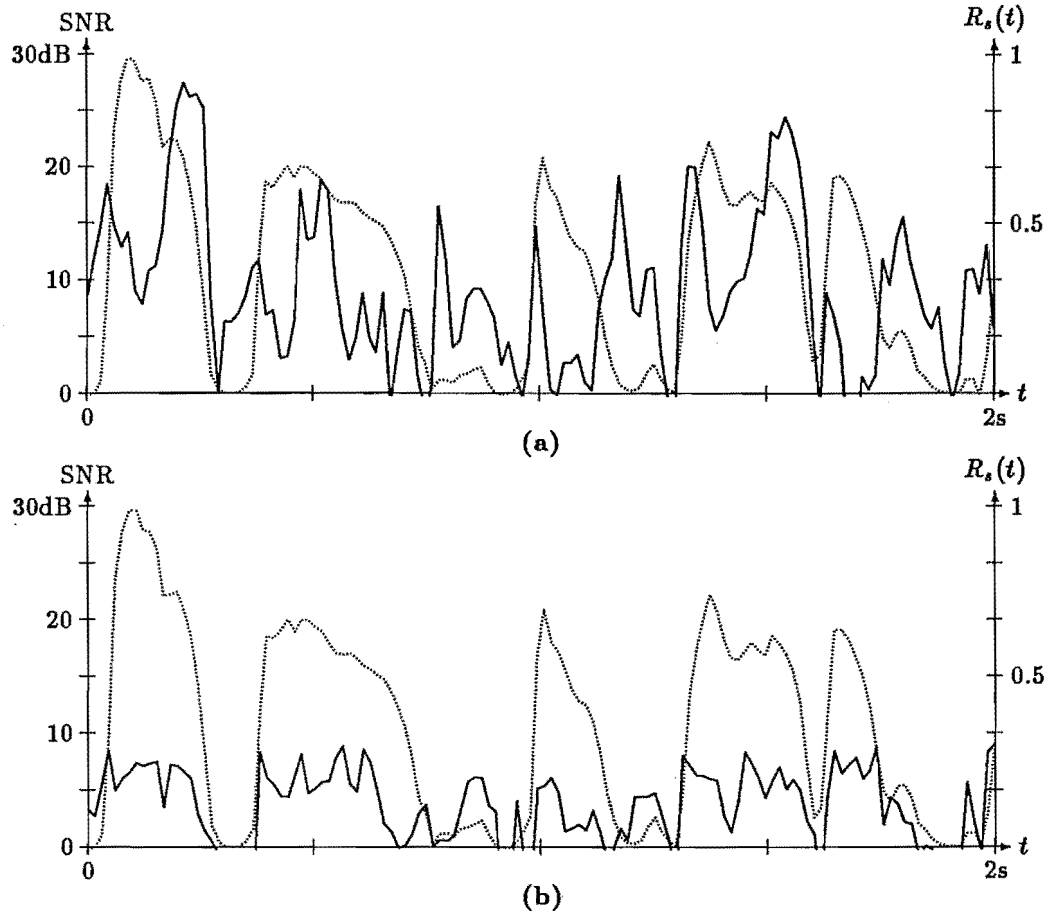


Figure 5.23. Segmental SNR (solid line) as a function of time during the first two seconds of **a:** the SAA/CLEAN processed utterance TC17KB (16.5kbit/s), and **b:** the MP-LPC processed utterance TM18KB (17.5kbit/s). The dotted line represents the RMS envelope of the speech signal.

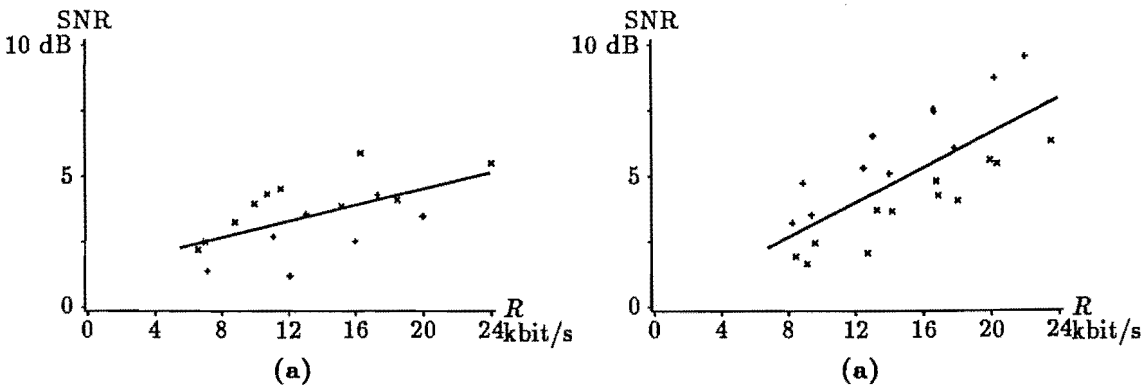


Figure 5.24. SNR as a function of data rate for utterances processed by **a:** SAA/CLEAN, and **b:** MP-LPC. The points marked with a + represent processed versions of the utterance TF-RAIN1, while those marked with a × refer to the utterance AM-RAIN1 (see Tables 5.2 and 5.3 for details). The line drawn on each graph is a linear regression through all the points.

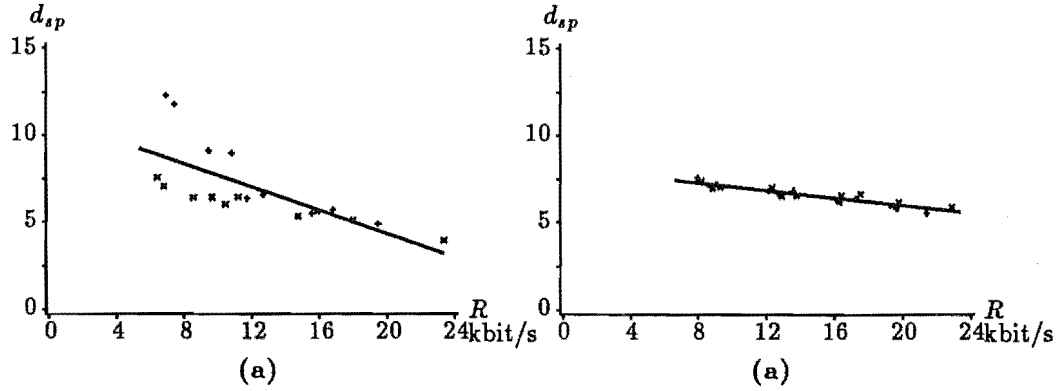


Figure 5.25. Log spectral distance measure (see §3.5.3.2) as a function of data rate for utterances processed by **a:** SAA/CLEAN, and **b:** MP-LPC. Refer to the caption to Fig.5.24 for further details.

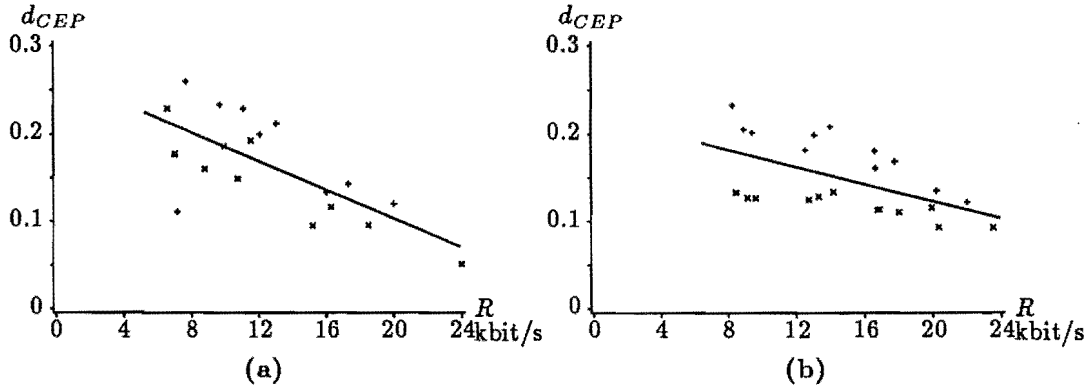


Figure 5.26. The cepstral distance measure of Kitawaki *et al.* (1988) (see §3.5.3.2) as a function of data rate for utterances processed by **a:** SAA/CLEAN, and **b:** MP-LPC. Refer to the caption to Fig.5.24 for further details.

The results presented here were obtained by means of the following procedure.

1. Two original utterances were employed, one each from a male and a female speaker (AM-RAIN1 and TF-RAIN1).

2. From each of the original utterances, the following test utterances were formed:

   - 4 reference utterances, corrupted with pink noise as described in §3.5.3.1. These had SNRs of 7,12,18, and 33 dB.
   - 9 utterances processed by SAA/CLEAN (as described in §5.3.2) at data rates ranging from 7kbit/s to 25kbit/s. See Table 5.2 for details.
   - 9 utterances processed by MP-LPC at data rates ranging from 8kbit/s to 26kbit/s. See Table 5.3 for details.

3. A menu-based speech play-back facility on an IBM PC-AT compatible computer was used to present the utterances. These were presented in four groups of randomly ordered utterances, labelled A1..A11, B1..B11, etc. The male and female utterances were in separate groups. Along with each group, a copy of the original utterance was also presented, labelled as such.

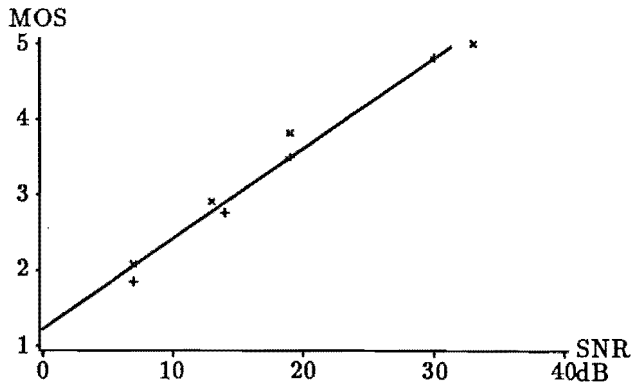


Figure 5.27. Mean opinion scores versus SNR for the reference utterances. + female and x male speakers.

4. Each subject (there were 17, comprising 10 males and 7 females) was given instructions to rate the quality of each utterance on a scale of 1 to 5, with the quality of the original utterance defined as 5. The subjects were provided with a descriptive interpretation of what each rating should be interpreted as (see §3.5.3.1). The subjects were told that they could listen to the utterances in any order and as many times as was necessary to judge the quality.

The scores assigned to the utterances were averaged across all 17 subjects to produce a "mean opinion score" (MOS) for each utterance. Fig.5.27 shows the relationship between MOS and SNR for the reference utterances. The MOS values for the utterances processed by MP-LPC and SAA/CLEAN respectively are shown as functions of the encoded data rate in Figs.5.28*a* and *b*. According to these results, there is little variation in the perceived quality of the processed utterances as the data rate is reduced down to about 10kbit/s. The quality drops off sharply below that rate, especially for the SAA/CLEAN encoded utterances. There does not appear to be any significant differences between the quality of the utterances spoken by male and female speakers. Fig.5.29 shows the variation in MOS values as the quantisation level is varied, at several different pulse rates. These curves indicate that the pulse amplitudes can be quantised fairly coarsely (q_{av} = 3.5 bits) without much effect on the reconstructed quality.

It is worth commenting, in a general way, on how well the MOS results reflect the variation in quality among the test utterances. Because categorical testing methods of the kind employed here limit subjects to a 5-point judgement scale, it is sometimes observed that such methods do not provide much discrimination between utterances of similar quality (cf. Rothauser *et al.*, 1971). Indeed, several of the subjects in this test mentioned that they scored many of the utterances as "3" even though they noticed differences between them — the differences were just too small to warrant a "2" or "4". This could partly explain why the MOS results flatten off at about 3. In retrospect, it could have been better to follow the suggestion of Rothauser *et al.* (1971) and use a 10-point scale. A preference type of test (§3.5.3.1) could also have been employed, but this demands much more of the subjects. Huggins and Nickerson (1985) suggest that preference and categorical methods produce similar results, so I did not consider the extra effort involved in performing a preference test to be warranted. Note, however, that the objective quality measures also level off at data rates above 16kbit/s, suggesting that the MOS results obtained are reasonable.

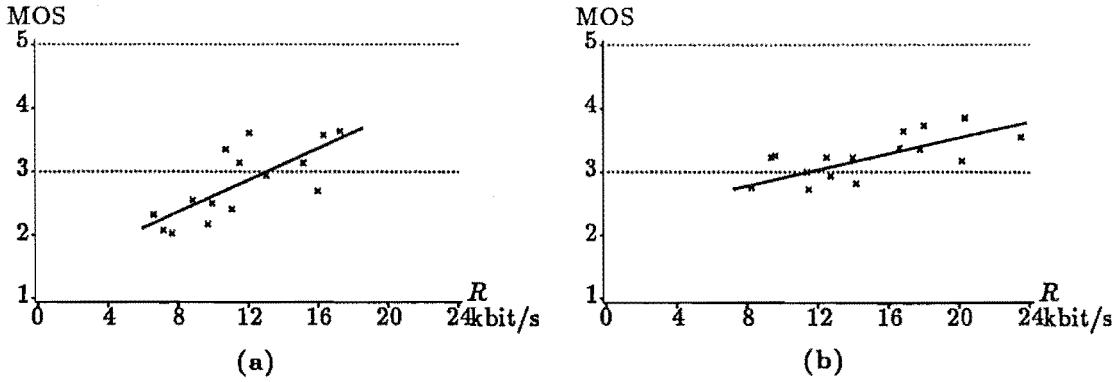


Figure 5.28. Mean opinion score versus data rate for the utterances that have been encoded by **a:** SAA/CLEAN, and **b:** MP-LPC. The superimposed line is a linear regression. Note that three of the utterances that had data rates greater than 24kbit/s (because of a high number of quantisation levels) are excluded from these graphs. They each had MOSs of about 3.5.

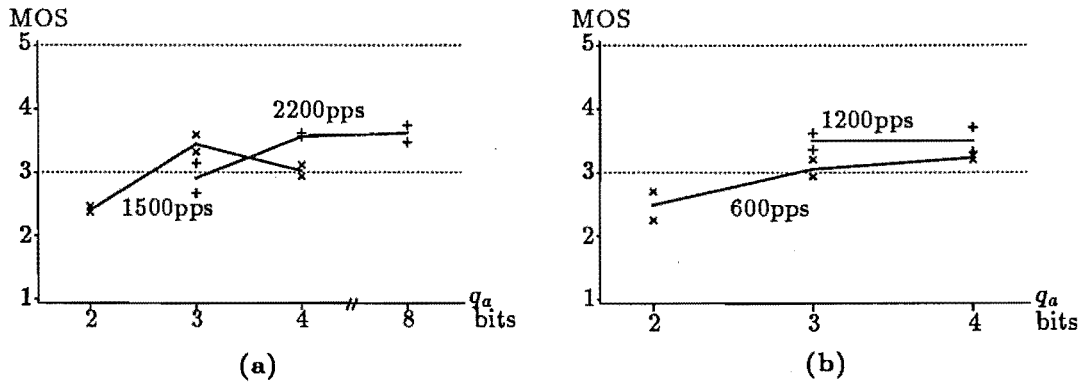


Figure 5.29. Mean opinion scores versus q_{av} — the quantisation level of the pulse amplitudes. Each curve represents the MOS for a different pulse rate, the rates being indicated on the figure. **a:** SAA/CLEAN, and **b:** MP-LPC.

5.3.3.3 Spectrograms of synthetic speech

The results presented in §5.3.3.1 and §5.3.3.2 indicate that SAA/CLEAN performs similarly to MP-LPC, although the quality of the reconstructed speech drops more rapidly at low data rates (<12kbit/s) for SAA/CLEAN than for MP-LPC. In order to examine more closely the form that this degradation takes, I here present spectrograms of (reconstructed) speech signals that have been encoded by the SAA/CLEAN and MP-LPC techniques.

Fig.5.30 shows the spectrogram of an unprocessed segment of the utterance AM-RAIN1. Spectrograms corresponding to the same segment after it has been encoded at various data rates by the SAA/CLEAN and MP-LPC techniques respectively are shown in Figs.5.31 and 5.32. Comparing the spectrograms of the processed speech signals with that of the unprocessed speech gives an indication the types of distortions that result from the encoding.

The spectrogram shown in Fig.5.31*a*, which is of an utterance that has been encoded by SAA/CLEAN at a data rate of less than 8kbit/s, appears "blurred", with indistinct formants. By contrast, the spectrogram of the 8kbit/s MP-LPC encoded

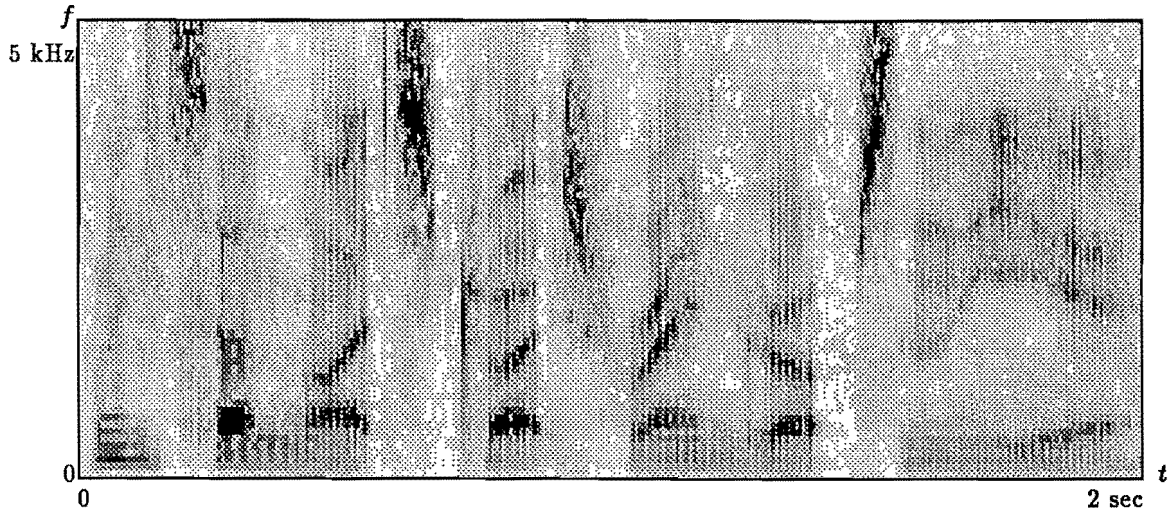


Figure 5.30. Spectrogram of the first two seconds of the utterance AM-RAIN1 (see Table 1.2 in §1.4.3).

utterance (Fig.5.32*a*) exhibits much more distinct formants. At higher data rates, the spectrograms of the SAA/CLEAN and MP-LPC encoded utterances are much more similar. In fact, at the higher data rates spectrograms of the MP-LPC encoded utterance can actually "look worse" than those of the SAA/CLEAN encoded utterance when compared with the original. This is probably due to the spectral weighting employed in the MP-LPC pulse optimisation, which increases the error at the formants (§3.5.2.4).

The more "distinct" formants produced by the MP-LPC technique, when compared with the SAA/CLEAN technique at low data rates, is to be expected, because the formants are well modelled by the LPC all-pole filter (which requires a data rate of only some 3–7kbit/s). The pulses in the MP-LPC technique are only required to provide the excitation energy and to "fill in the details" not modelled by the LPC all-pole filter. By contrast, the CLEAN pulses are required to model both the excitation sequence (although the excitation "shaping" is accomplished by the SAA signal), and the formant structure of the speech signal. With only a few pulses, SAA/CLEAN can only approximate the formant structure (see §5.4.3). Furthermore, because the sequence of CLEAN pulses represents an all-zero rather than an all-pole filter (as modelled by the LPC coefficients), more pulses are required to match the resonant characteristics of the formants. Further discussion on the differences, similarities, and relative advantages of the MP-LPC and SAA/CLEAN methods of speech coding appears in §5.4.2.

5.4 Discussion: Approaches to interpreting SAA/CLEAN analysis of speech

In this section I draw together some of the points made elsewhere in Chapters 4 and 5 about the meaning of the SAA and CLEAN signals. §5.4.1 discusses the relationship of this speech model to the source-filter model of speech, while §5.4.2 makes some comments on the similarities and differences between the CLEAN and MP-LPC techniques. Finally, §5.4.3 describes the somewhat different view of SAA and CLEAN as seen from the frequency domain.

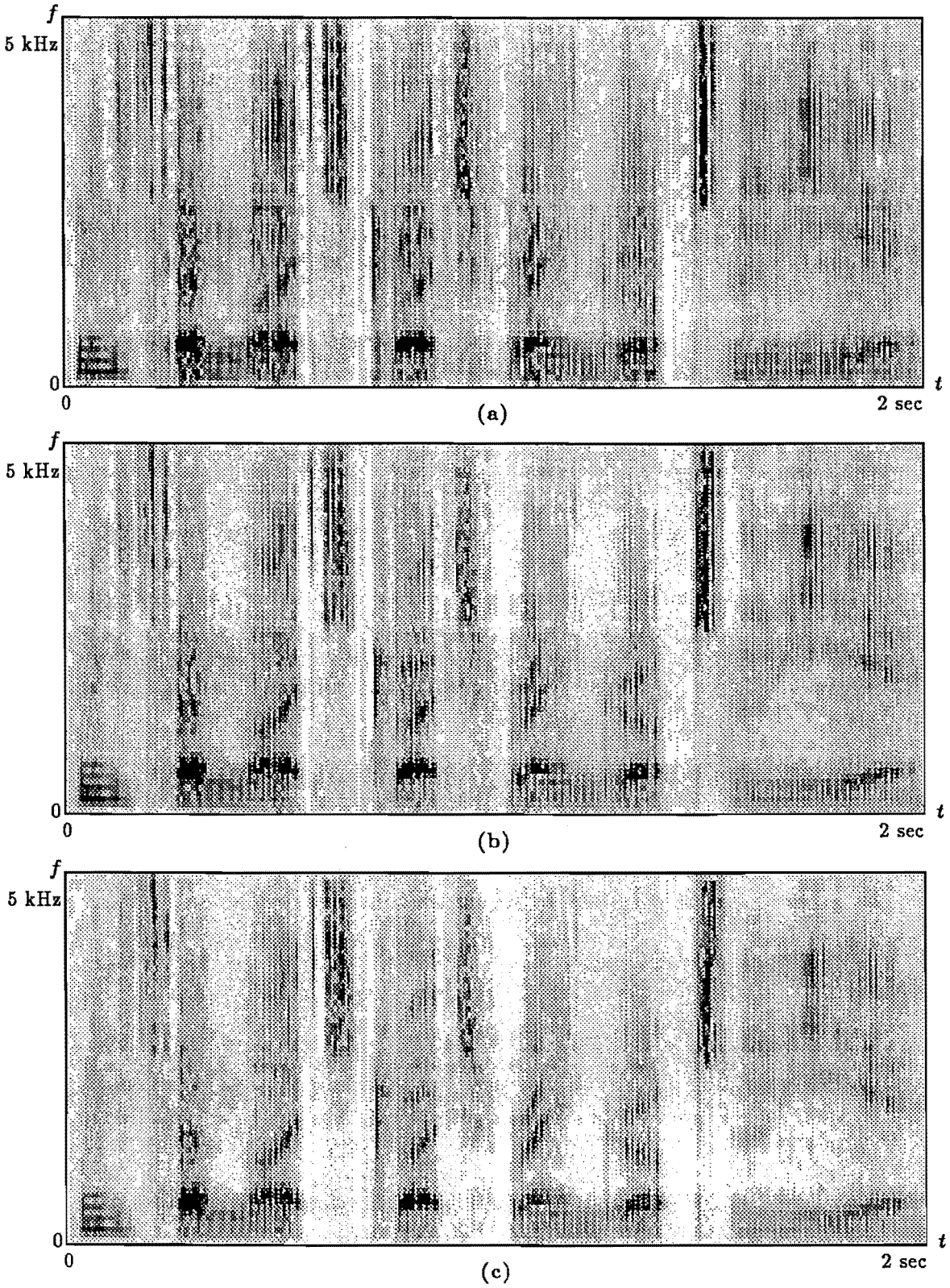


Figure 5.31. Spectrograms of an utterance (AM-RAIN1) that has been encoded by SAA/CLEAN. Utterances (see Table 5.2) **a:** AC7KB, **b:** AC12KB, and **c:** AC16KB.

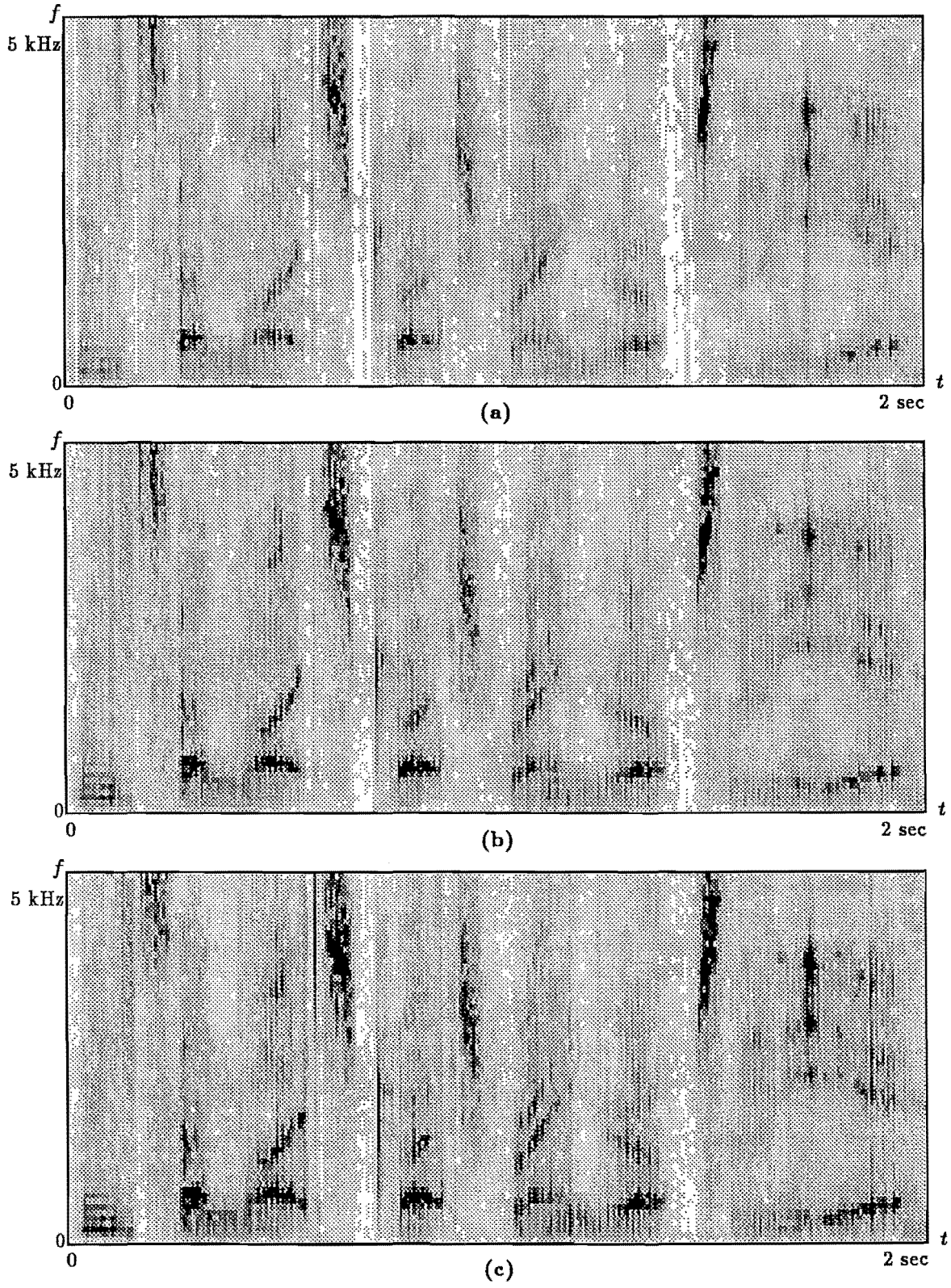


Figure 5.32. Spectrograms of an utterance (AM-RAIN1) that has been encoded by MP-LPC (see §3.5.2.4). Utterances (see Table 5.3) **a:** AM8KB, **b:** AM14KB, and **c:** AM18KB.

5.4.1 Relating the CLEAN signal to the vocal tract filter

Shift-and-add of an utterance produces a signal that represents the long-term "invariant" characteristics of the utterance. §4.3 presents results which indicate that the SAA signal can be roughly associated with the average glottal excitation — but with a not insignificant component due to the long-term "invariant" vocal tract response. The CLEAN signal, obtained as described in this chapter, plays the complementary role of representing the "dynamic" aspects of an utterance. As a counterpart to §4.3 then, it seems appropriate to elucidate the extent to which the CLEAN signal can be associated with the vocal tract filter response. §5.4.1.1 recapitulates the variant/invariant model of speech introduced in §4.2.1, while §5.4.1.2 discusses, in a general way, the variant and invariant components of the source-filter speech model. Finally, §5.4.1.3 discusses one way in which the CLEAN signal can be directly associated with the vocal tract filter.

5.4.1.1 Invariant/variant decomposition of speech

Recall from §4.2.1 the model invoked to introduce SAA processing of speech. This model represents a speech signal $s(t)$ as

$$s(t) = \sum_{m=1}^{M} s_m(t - T_m), \tag{5.28}$$

where each short segment $s_m(t)$ is separated into three components according to the expression

$$s_m(t) = s^i(t) \odot s^v_m(t) + s^c_m(t), \tag{5.29}$$

with $s^i(t)$, $s^v_m(t)$, and $s^c_m(t)$ being called the invariant, convolutional variant, and additive variant (contamination) terms respectively. The form of $s^i(t)$ is fixed by SAA to be

$$s^i(t) = \langle s_m(t) \rangle_m, \tag{5.30}$$

since T_m in (5.28) is defined as the instant where $|s_m(t)|$ is greatest. Fixing $s^i(t)$ in this way implies that $s^v_m(t)$ and $s^c_m(t)$ both have negligible ensemble averages, but, because of the general inconsistency of convolution (§1.2.5.3), there are an infinite number of different signals $s^v_m(t)$ and $s^c_m(t)$ that satisfy (5.29). The CLEAN method of deconvolution described in §5.2 finds a solution that satisfies the restrictions

- $s^v_m(t) = \sum_{k=1}^{N_p} v_k \delta(t - p_k)$ [where N_p is small with respect to the Nyquist sampling requirements (see §5.2.5.4 and §5.4.3.4) (5.31)
- $\langle s^c_m(t)^2 \rangle_t$ is minimised (§5.2.6). (5.32)

Note that the two restrictions (5.31) and (5.32) may be incompatible for some segments of speech. §5.4.1.2 discusses the general characteristics of speech signals which allow one to make assumptions about the relative importance of $s^v_m(t)$ and $s^c_m(t)$. §5.4.3.2 also discusses this incompatibility, with regard to instability in the CLEAN algorithm.

5.4.1.2 Speech components — Variant and invariant; source and filter

The traditional simplified model of speech is the source-filter model, which represents a speech signal by the convolution between two physiologically relevant quantities, the excitation "source" and the vocal tract "filter" (§2.3.1.4). By contrast, the model of speech outlined in §5.4.1.1 represents a speech signal by means of components which are

defined in terms of the long-term behaviour of the speech signal, namely, "invariant" and "variant" components. The results presented in §4.3 suggest that the invariant component $s^i(t)$ of the latter model contains contributions from the long-term averages of both the source and filter components of the source-filter model. The variant components $s^v_m(t)$ and $s^c_m(t)$ are thus formed by the variations of both the source and filter components throughout an utterance.

For the SAA/CLEAN technique to successfully represent speech signals, it is important that $s^c_m(t)$ be in some way small relative to $s^v_m(t)$ — because $s^c_m(t)$ is discarded as a residual "error" during the CLEAN processing. In this section I discuss, in qualitative terms, the variant and invariant characteristics of the source and filter speech components, thereby reaching some conclusions about the relative importance of $s^v_m(t)$ and $s^c_m(t)$ in representing $s_m(t)$.

First, it is useful to recall the source-filter model of speech production (§2.3.1.4). A voiced speech signal $s(t)$ is represented by the convolution between a quasi-periodic "pitch train"

$$g_p(t) = \sum_k \delta(t - T_k) \tag{5.33}$$

of unit impulses, the k^{th} of which is positioned at $t = T_k$; a glottal shaping filter $g_k(t)$; and a vocal tract filter $v_k(t)$. The speech signal is therefore described by the equation

$$s(t) = \sum_k g_k(t - T_k) \odot v_k(t). \tag{5.34}$$

The source-filter model as expressed by (5.34) can be equated to the variant/invariant model of (5.28) and (5.29) by separating both $g_k(t)$ and $v_k(t)$ into variant and invariant components, substituting them in (5.34), and associating the invariant component of the expanded expression with $s^i(t)$ and the variant components with $s^v_m(t)$ and $s^c_m(t)$ (see §4.2.1 for the details). For the reader's convenience, I repeat the definitions developed in §4.2.1:

$$s^i(t) = g^i(t) \odot v^i(t), \tag{5.35}$$

$$s^v_m(t) = g^v_m(t) \odot v^v_m(t), \tag{5.36}$$

and

$$s^c_m(t) = v^c_m(t) \odot g^i(t) \odot g^v_m(t) + g^c_m(t) \odot [v^c_m(t) + v^i(t) \odot v^v_m(t)], \tag{5.37}$$

where the superscripts i, v, and c imply the invariant, variant, and "contamination" terms respectively of the expanded signal quantities.

The glottal pulse, while it varies due to factors such as vocal intensity and pitch, can usefully be considered as invariant, as is demonstrated by the ability of speech synthesisers employing a fixed pulse shape to synthesise reasonable quality speech (§3.5.2). Physically, the variations between each pulse are induced by the changes in the mechanical properties of the vocal cords that result from muscular action (cf. §2.3.1.1; §2.2.1.2). The variations are modelled by a convolutional component $g^v_m(t)$ and an additive component $g^c_m(t)$. Because each glottal pulse is a separate "event", there is no physical implementation of a "convolution" between $g^i(t)$ and $g^v_m(t)$. Hence it is likely that the additive term $g^c_m(t)$ is as significant as $g^v_m(t)$ in describing the variation in pulse shape (recognising the point made in §1.2.5.3 that there are an infinite number of signal pairs $g^v_m(t)$ and $g^c_m(t)$ that satisfy the inconsistent convolution).

The variable components of the vocal tract impulse response are intuitively much greater than those of the glottal pulse, since it is largely by such changes that a person utters different speech sounds (cf. §2.2.1). In fact, if the vocal tract varied in a completely unbiased manner, so that its average impulse response was negligible, the invariant component $v^i(t)$ (defined by (4.10) in §4.2.1) would become an impulse

function. Consequently, $v_m(t)$ could then be completely described by the component $v_m^v(t)$, and the additive term $v_m^c(t)$ could be discarded. In practice (see the results in §4.3.2), the average vocal tract impulse response $v^i(t)$ is likely to be non-impulsive due to the limited mobility of the articulatory organs, the particular style of pronunciation of the speaker, and any phonetic imbalance in the utterance. If (despite the last sentence) $v^i(t)$ is assumed to be "almost" impulsive, one can infer that $v_m^v(t)$ is more significant than $v_m^c(t)$ for representing the variation of $v_m(t)$ from segment to segment.

From the discussion in the previous two paragraphs, it can be seen that the relative significance of the "contamination" component $s_m^c(t)$ in any segment depends mainly on how large $g_m^c(t)$ is. For segments in which the shape of the excitation pulse $g_m(t)$ is very different from $g^i(t)$, and when the differences cannot be modelled by the convolutional component $g_m^v(t)$, the definitions (5.36) and (5.37) imply that the contamination term $s_m^c(t)$ can, for some values of m, even dominate $s_m^v(t)$. §5.4.3.2 contains further discussion of this point in terms of the frequency domain representation of the CLEAN algorithm.

5.4.1.3 Physical interpretation of the CLEAN signal

In §4.3 the SAA signal is shown to approximately represent the average glottal excitation $g(t)$ of an utterance, together with a component due to the average vocal tract impulse response. This additional component cannot be easily removed from the SAA signal (see §4.3.3 for one approach), and its contribution depends upon the characteristics of both the speaker's talking style and the utterance. Assuming, however, that the component due to the average vocal tract impulse response $v^i(t)$ can be ignored, the SAA signal can be taken to approximate $g^i(t)$. By letting $g_m^v(t) = A_m\delta(t)$, where A_m is the amplitude of the m^{th} occurrence of $g^i(t)$, the CLEAN signal becomes $v_m^v(t)$, which approximately represents the vocal tract filter $v_m(t)$.

Now that I have managed to associate the CLEAN signal with the vocal tract impulse response — by the simple expedient of discarding all the other terms as negligible (!), I can proceed to describe one way in which the CLEAN signal can be physically interpreted. Recall that the CLEAN signal consists of only a few non-zero "pulses". Such a signal could represent the impulse response of a uniform-tube model of the vocal tract (§2.3.1.3), with each pulse representing the superposition of all the multiply reflected impulses that arrive at the lips at a particular instant.

In the usual formulation of the uniform-tube vocal tract model (cf. Wakita, 1973), all segments are considered to be of the same length, corresponding to an acoustic time delay of one sampling interval. Thus an output "pulse" occurs at every sampling instant, and the impulse response is effectively smooth. Furthermore, because each segment is of the same length, it is relatively straightforward to compute the tube segment areas (assuming suitable boundary conditions) from the impulse response at the lips or even the speech signal itself (cf. Sondhi and Gopinath, 1971; Wakita, 1973).

If the segments of the uniform-tube vocal tract model are allowed to be of arbitrary length, such that the acoustic delays through each segment are longer than the sampling interval, the resulting impulse response exhibits discrete pulses separated by intervals when no output signal occurs. Each pulse represents the superposition of one or more multiply reflected source impulses. However, calculating the dimensions of the tube segments from the pulse sequence is a difficult problem and is beyond the scope of this thesis (see Sondhi, 1984 and Millane and Bates, 1982 for some background on related problems).

It is important to note that the usual formulation of *reflection coefficients* (§3.2.3) is in terms of an all-pole model, while the CLEAN pulses described here represent the impulse response (or all-zero formulation) of the reflections. Hence the

standard reflection coefficients represent the actual reflections that occur at each discontinuity within the resonant cavity. Each CLEAN pulse, on the other hand, represents the superposition of all the multiply reflected copies of the input pulse that leave the mouth at a particular instant.

5.4.2 Comparison with multi-pulse LPC

Several other sections of this chapter remark on the similarities between the CLEAN deconvolution and the MP-LPC pulse estimation schemes. Here I discuss the points of similarity and difference between the two techniques, explaining how the particular ways in which each scheme models the speech signal affects its performance.

Both the MP-LPC and SAA/CLEAN approaches to speech modelling represent the speech signal by a convolution between a sparse pulse sequence and a shaping filter. However, both the pulse sequence and filter are obtained by different methods, and model different aspects of the speech signal, in the two schemes. §5.4.2.1 details the differences and similarities in the techniques that are invoked to obtain the pulse sequence and filter components, while §5.4.2.2 discusses the components in terms of the information in the speech signal that they represent.

5.4.2.1 Differences and similarities in the analysis techniques

The "filter" in the SAA/CLEAN technique represents the long-term average, or invariant, component of the speech signal (§5.4.1). It is computed over intervals of some 5–10s by means of the SAA technique described at length in Chapter 4. In the MP-LPC technique, the "filter" is actually a sequence of different filters, each representing the waveform of the speech signal during a short interval (10–20ms). These filters are formulated in terms of an all-pole model with a few (typically 8–16, see Kroon and Deprettere, 1988) coefficients. They are computed by means of the LPC algorithm, which minimises the least-squares error between the speech waveform and the filter's impulse response (§3.2).

The SAA/CLEAN pulse sequence is obtained by means of the CLEAN subtractive deconvolution algorithm (this chapter). This algorithm iteratively locates pulses at the position at which the magnitude of the dirty signal (which is initially set equal to the speech signal) is at a maximum. The new pulse amplitude is set equal to a proportion of this maximum, and the pulse is added to the "CLEAN" signal. A scaled copy of the SAA signal (representing the filtered pulse) is then subtracted from the dirty signal for the next iteration. The iterations are continued until either the average level of the dirty signal has been sufficiently reduced, or until enough (as discussed in §5.2.5.4) pulses have been located. After this, the amplitudes of the pulses are re-optimised in a least-squares fashion.

There are several techniques for computing the MP-LPC pulse sequence (cf. Singhal and Atal, 1989; Kroon and Deprettere, 1988). Probably the simplest is the method described in §3.5.2.4, in which at each iteration, a new pulse is placed at the instant where the magnitude of the cross-correlation between the filter impulse response and the "dirty signal" (where I use this term to emphasise the similarities with the CLEAN algorithm) is greatest. Unlike the CLEAN algorithm, subsequent iterations do not usually consider the positions of previously located pulses as candidates for new pulses. The amplitude of the new pulse is set to a proportion of the maximum value of the cross-correlation. Note that the cross-correlation signal does not need to be re-computed at each iteration because it can be updated from its value at the previous iteration. Generally a fixed number of iterations are performed, resulting in a specified

number of pulses for each segment. The amplitudes of the pulses are often re-optimised in the same way as the CLEAN pulses.

The brief descriptions in the previous two paragraphs indicate the close similarities between the MP-LPC and CLEAN pulse searching algorithms. In fact, the main difference is that the pulses are positioned at the maximum magnitude of the dirty signal in the CLEAN algorithm, but at the maximum magnitude of the cross-correlation between the dirty signal and the filter impulse response in the MP-LPC algorithm. §8.2.2.2 describes how the CLEAN algorithm can be modified to locate pulses using this cross-correlation approach. It turns out that the MP-LPC algorithm is equivalent to performing CLEAN on the cross-correlation between the speech signal and the filter impulse response, with the auto-correlation of the filter impulse response as the CLEAN kernel (see §8.2.2.2).

Note that another difference between the CLEAN and MP-LPC pulse searching algorithms is that MP-LPC usually includes some spectral weighting of the filter (Atal and Remde, 1982) so that the error residual is concentrated under the speech formants, where it is of less perceptual importance (cf. Schroeder *et al.*, 1979). It is not clear how to incorporate some such spectral weighting into the CLEAN algorithm, because the speech spectrum is not directly available as it is in MP-LPC (the LPC coefficients can be manipulated as if they refer to the speech spectrum).

Because MP-LPC models the vocal tract filter with a time-varying all-pole filter, there is no need to separate the voiced and unvoiced parts of speech as is done in the SAA/CLEAN technique. In addition, the LPC filter matches the speech spectrum fairly closely within each segment, so fewer pulses are required for accurate reconstruction than with SAA/CLEAN. However, the LPC parameters must also be stored or transmitted, which in turn increases the data rate. The LPC parameters can be encoded efficiently with 36 bits required for each set of 10 coefficients (Kroon and Deprettere, 1988). Hence 1800 bits/s are required if the filter is updated at 50 Hz, and 7200 bits/s are required at an update rate of 200 Hz.

In terms of the computational requirements, the two methods are roughly similar. Additional computation is required in the MP-LPC technique because of the need to extract the LPC coefficients, encode them for transmission, and compute their impulse response for use in the pulse searching algorithm. However, the SAA/CLEAN technique described here requires that the speech signal be filtered into two sub-bands, with all subsequent operations being performed in parallel on both sub-bands. This actually increases the computational load by less than twice, since fewer pulses are found in the high frequency sub-band and because the high frequency CLEAN kernel can be made shorter than the low frequency one.

5.4.2.2 Interpretation of pulse sequence and filter components

In terms of information in speech, the LPC filter can be associated with the vocal tract filter, with a direct relationship between the filter coefficients and the shape of the vocal tract (§3.2.3). The LPC coefficients have been succesfully employed in speech recognition schemes (§3.6.1), which implies that they efficiently represent the linguistic information within an utterance. The MP-LPC pulse sequence attempts to make the speech reconstructed from the LPC coefficients sound more "natural". It can thus be thought of as an ancillary signal which represents those parts of a speech signal that are not modelled by the all-pole LPC filter. Note that the multi-pulse sequence was originally developed as an improved excitation source for LPC vocoders, producing more natural sounding speech, and implicitly representing both the voiced/unvoiced and pitch information which is otherwise necessary (see §3.5.2.3).

In the SAA/CLEAN technique, the pulse sequence represents aspects of both

the excitation sequence (i.e. the pitch periodicity for voiced speech) and the vocal tract filter. In order to utilise the linguistic information inherent within the pulse sequence, the pulses within each pitch period must be identified (see §5.4.1.3). Further research is obviously required to ascertain whether the CLEAN pulses can be employed in a recognition scheme as useful descriptors of the linguistic content of speech (see §8.2.2).

Note that the LPC filter is an all-pole filter, which efficiently matches the series of resonators that comprise the vocal tract filter. By contrast, the CLEAN signal during each pitch period is an all-zero representation of the "vocal tract filter". It is therefore not so adept at representing sharp resonances with only a few non-zero coefficients. Further discussion of this point appears in §5.4.3.1.

5.4.3 CLEAN in the frequency domain

As described in §2.1.4, speech sounds are characterised by their formants, which are spectral peaks representing the resonances in the vocal tract. Because of the important role that spectral representations of speech play in the traditional descriptions of speech sounds and analysis methods (e.g. §2.1.4, §3.3, §3.5), it is useful to examine the CLEAN algorithm from this point of view. This section describes the CLEAN algorithm in terms of its effects on the speech spectrum. §5.4.3.1 presents examples of spectra of CLEAN signals to illustrate the manner in which SAA/CLEAN analysis models the speech spectrum. §5.4.3.2 then discusses the convergence difficulties of the CLEAN algorithm from the point of view of spectral inconsistency between the SAA and speech signals. §5.4.3.3 indicates how this instability is partially allayed by processing the speech in two sub-bands. Finally, §5.4.3.4 introduces the concept of non-uniform sampling as an approach to understanding SAA/CLEAN speech analysis.

5.4.3.1 A spectral view of the CLEAN signal

Recall that in the radio-astronomical application of CLEAN (§5.1.1; Högbom, 1974; Thompson *et al.*, 1986), Fourier space contains large gaps where measurements are not made. Schwartz (1978) shows that CLEAN effectively performs a least-squares interpolation between the non-zero Fourier samples. Speech spectra are somewhat different than radio-astronomical spectra (!), but they are typically of large dynamic range (see Fig.4.29 in §4.3.1.2). CLEAN can therefore be viewed as a type of spectral flattening. As a deconvolution technique, it removes the severe low-pass filtering effect of the glottal shaping filter (for the voiced sections of speech. The general situation is treated in §5.4.3.3).

Fig.5.33 shows the time domain and log spectrum of a segment of the utterance AM-RAIN1. CLEAN signals obtained from this segment, having pulse rates ranging from 500pps to 2000pps, are shown in Figs.5.34*a* through *d*, while their log spectra appear in Fig.5.35. Note how the spectra of the CLEAN signals are very "rippled". In the frequency domain, the invariant/variant speech model (§5.4.1.1) is expressed as

$$S_m(f) = S^i(f)S_m^v(f) + S_m^c(f), \tag{5.38}$$

where the upper case quantities are the Fourier transforms of the respective lower case quantities in (5.29). By replacing $S_m^v(f)$ with the ensemble of CLEAN pulses $\{v_k; p_k\}$, (5.38) can be written as

$$S_m(f) = S^i(f)\sum_{k=1}^{N_p} v_k e^{-i2\pi f p_k} + S_m^c(f). \tag{5.39}$$

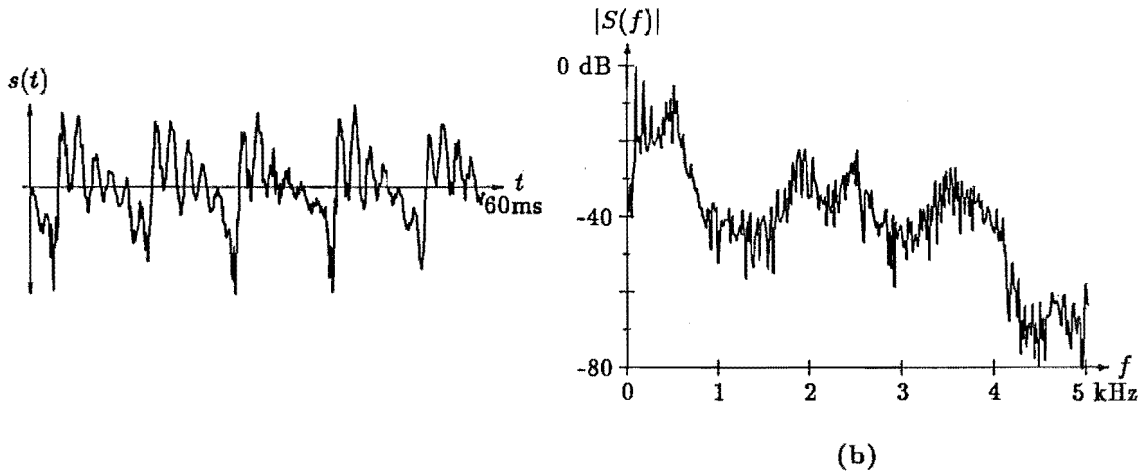


Figure 5.33. a: Time domain and b: log spectrum of a segment of the utterance AM-RAIN1 (the dʒ in "raindrops"). Note that the vertical axis of the graph in *b* has a range of 80dB rather than 60dB as is usual in the remainder of this thesis.

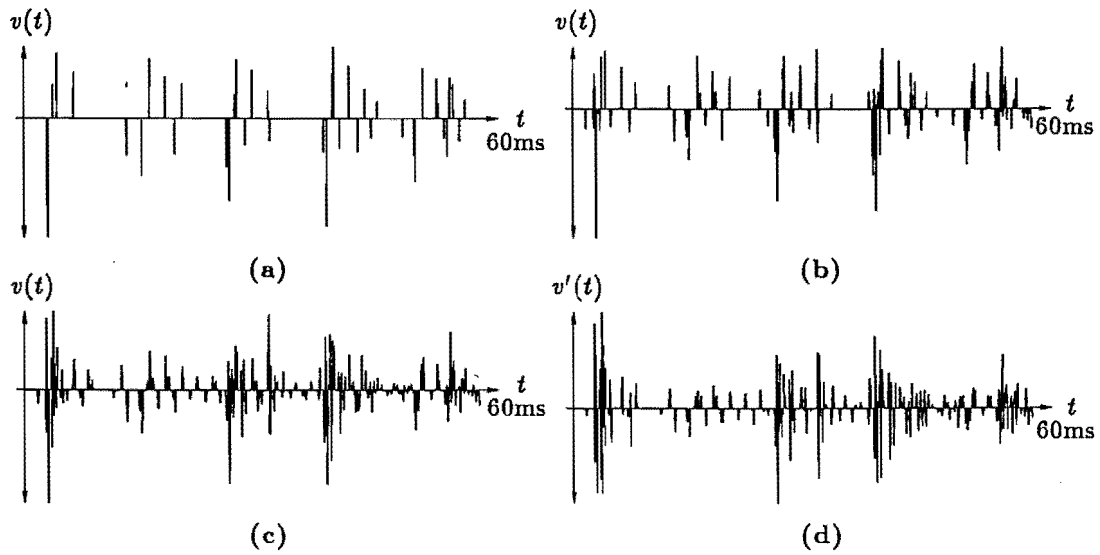


Figure 5.34. CLEAN signals obtained from the segment of speech shown in Fig.5.33*a*. Pulse rates of a: 500pps, b: 1000pps, and c: 2000pps. d: The optimised version of the CLEAN signal shown in c. The SNRs of the reconstructions of these four CLEAN signals are 8, 11, 15, and 18dB respectively.

This equation asserts that the spectrum of the CLEAN signal consists of a sum of sinusoidal oscillations, each at a "frequency" proportional to the position of one of the pulses and weighted by the magnitude of that pulse (Schwartz, 1978). The reconstructed signal is obtained by multiplying the combined sinusoids with the spectrum of the SAA signal.

As the number of pulses increases, the sinusoidal "ripples" reinforce in those parts of the speech spectrum where the formant peaks exist. Note that the sinusoids that comprise the CLEAN spectrum in (5.39) extend over the entire range of the Fourier domain. They therefore extend into frequencies where $S^i(f)$ is negligible. Although the magnitude of the components of the CLEAN spectrum are arbitrary in those parts

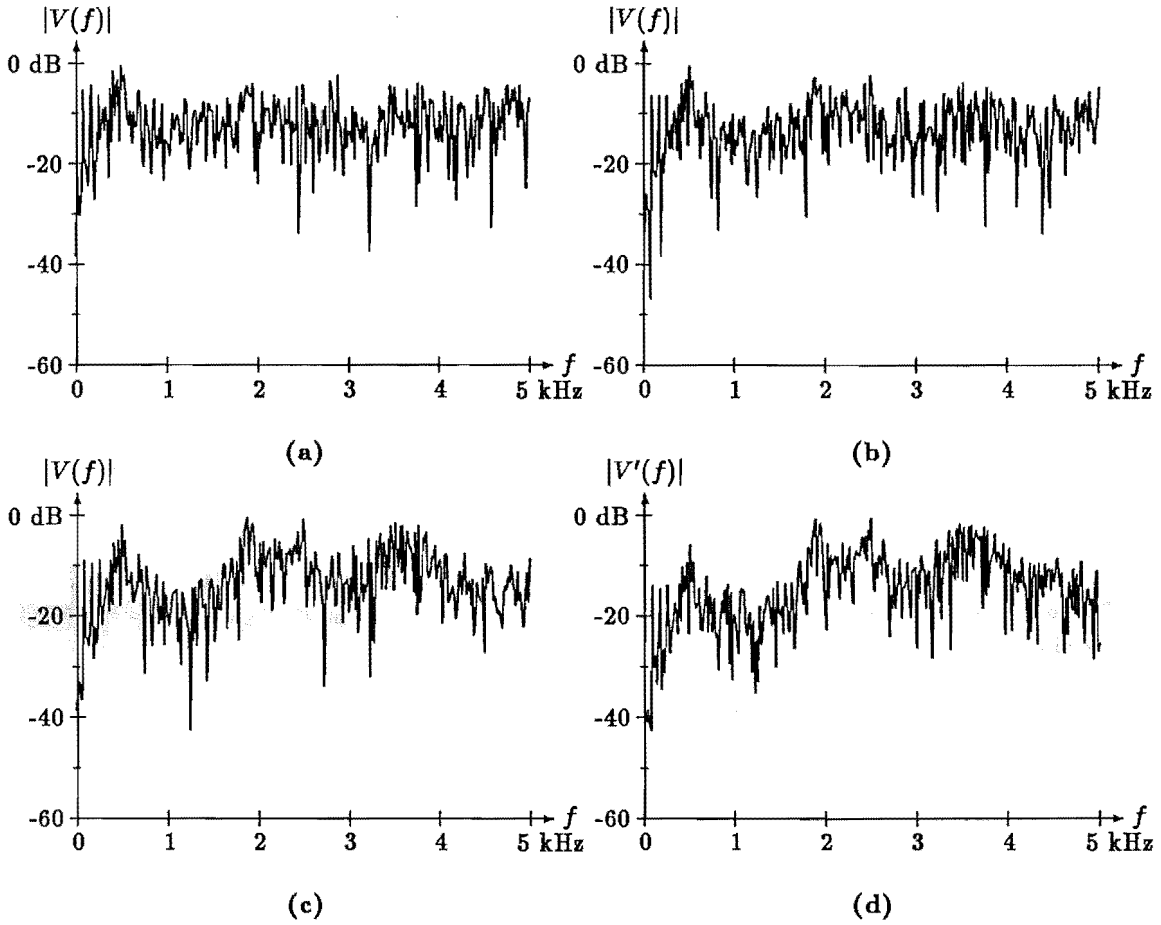


Figure 5.35. Log spectra of the CLEAN signals shown in Fig.5.34.

of the Fourier domain where $S^i(f)$ is negligible, they do not become unmanageably large as occurs in multiplicative deconvolution. This is because all the "deconvolution" operations in the CLEAN algorithm are performed on the dirty signal, which has effectively been filtered by $S^i(f)$ so that it does not contain energy at the frequencies where $S^i(f)$ is negligible (that is if the additive contamination is negligible. §5.4.3.2 discusses the effects of significant contamination). Each pulse is therefore positioned so as to reduce the error within the frequency bands where $S^i(f)$ is significant. The superposition of the sinusoids outside these frequencies tends to be relatively "flat". It is in this sense that CLEAN can be called a spectral flattening algorithm.

5.4.3.2 Instability in CLEAN

The CLEAN signal is represented in the frequency domain by (5.39). In terms of the notation employed in §5.2.3 to describe the CLEAN algorithm, $S^i(f)$ becomes $G(f)$, the spectrum of the CLEAN kernel $g(t)$, and $S^c_m(f)$ becomes $R_m(f)$, the spectrum of the residual (error) signal $r_m(t)$.

As discussed in §5.4.3.1, the speech spectrum $S_m(f)$ is composed of the superposition of N_p scaled and phase-shifted copies of $G(f)$, together with an additive error $R_m(f)$. The presence of the additive contamination implies that if $G(f)$ is of negligible amplitude at some frequency at which the amplitude of $S_m(f)$ is not negligible, the amplitude of $R_m(f)$ at that frequency cannot be reduced by any number of CLEAN pulses. Furthermore, if $G(f)$ contains energy at frequencies where $S_m(f)$ does not, subtracting the kernel from the speech signal during the CLEAN procedure (§5.2.3)

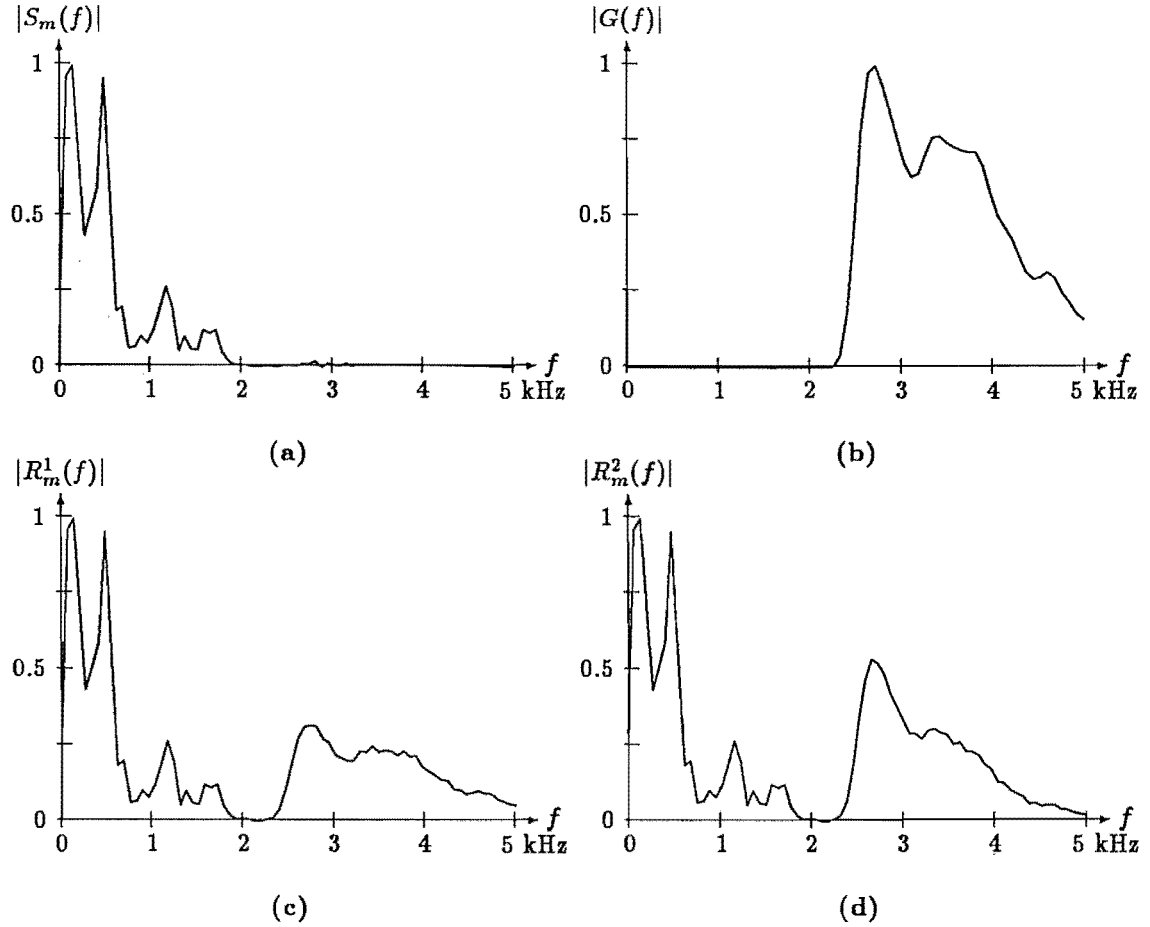


Figure 5.36. Frequency domain illustration of the instability that occurs in CLEAN when the dirty signal and kernel are inconsistent. **a:** Original spectrum $S_m(f)$. **b:** Spectrum of kernel $G(f)$. **c:** Residual spectrum after one iteration, and **d:** after two iterations.

actually adds energy to $R_m(f)$ at that frequency — which must then be removed by a subsequent subtraction. Thus the CLEAN algorithm can become unstable if $S_m(f)$ and $G(f)$ are (cf. §1.2.5.3). This is illustrated by the extreme example shown in Fig.5.36. The spectrum $S_m(f)$ is shown in Fig.5.36*a*, while Fig.5.36*b* shows $G(f)$. After one iteration ($N_p = 1$), $R_m(f)$ is as shown in Fig.5.36*c*. As shown, subtracting the kernel has not only failed to remove any of the energy in $S_m(f)$, but it has added energy that was not there previously. The residual spectrum $R_m(f)$ after a subsequent iteration is shown in Fig.5.36*d*. The time domain versions of these signals are shown in Fig.5.37.

In terms of the speech model outlined in §5.4.1, the instability described above arises when the additive component $g_m^c(t)$ is large relative to $g^i(t)$ and $g_m^v(t)$ (see the final paragraph of §5.4.1.2). A large $g_m^c(t)$ implies that $G_m(f)$ (and hence $S_m(f)$)contains significant energy at frequencies where $G^i(f)$ is negligible.

Note that the instability difficulties described above do not occur in the radio-astronomical application of CLEAN (see §5.1.1). In radio-astronomical CLEAN, the kernel is consistent with the dirty signal by definition, because both spectra are non-zero at the same points (and only those points) in Fourier space. The kernel used for CLEANing of speech signals, however, is a long-term average, so may be inconsistent with some particular segments of the speech signal.

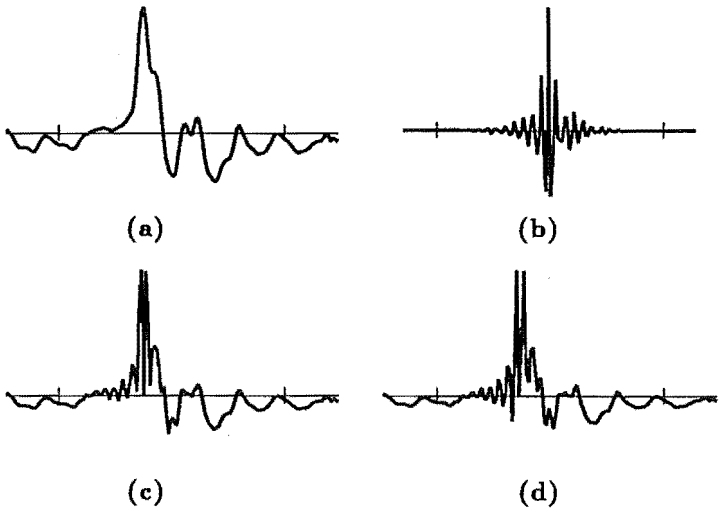


Figure 5.37. Instability in CLEAN. Time domain versions of the spectra shown in Fig.5.36. **a:** $s_m(t)$, **b:** $g(t)$, **c:** $r_m^1(t)$, and **d:** $r_m^2(t)$.

5.4.3.3 Separation into spectral sub-bands

Because of the disparate spectral characteristics of the voiced and unvoiced parts of speech, the same "spectral flattening filter" cannot be applied to both parts (see §5.4.3.2). In order to deal adequately with these differences, I separate speech signals into low and high frequency sub-bands before performing SAA and CLEAN (§4.2.4.5; §5.2.5.5). The frequency domain view of this action is shown in Fig.5.38. This figure shows the spectra of SAA signals computed from the low and high frequency sub-bands of the utterance AM-RAIN1, together with the spectrum of the SAA signal computed from the entire utterance (Fig.5.38*c*). Comparing Figs.5.38*a* and *b* with Fig.5.38*c* indicates that the SAA signal of the high frequency sub-band appears to represent the energy in that band better than does the SAA signal obtained from the entire signal, which has negligible energy in its high frequency components.

The spectra of the low frequency and high frequency sub-bands of a segment of speech are shown in Fig.5.39, while those of the respective CLEAN signals appear in Fig.5.40. Notice that the CLEAN spectra extend over the half of the Fourier domain that is zero in the filtered versions of $S^i(f)$ and $S_m(f)$. This "spectral interpolation" corresponds to the spectral flattening property of CLEAN described in §5.4.3.1. However, this "interpolation" into the other sub-band is subsequently removed when the CLEAN signal is filtered with the SAA signal to form the synthetic reconstruction. Spectra of the reconstructions of the low and high frequency sub-bands are shown in Figs.5.41*a* and *b* respectively. The reconstruction of the entire signal, formed by adding together the two sub-band reconstructions, is depicted in Fig.5.42*a*. For comparison, Fig.5.42*b* shows the spectrum of the reconstruction formed from the CLEAN signal of the entire signal (as depicted in Fig.5.35*c*).

Performing CLEAN on the two sub-bands also allows one to take advantage of the lower requirements for fidelity in the higher frequency band by employing fewer pulses and coarser quantisation levels to encode that band. This results in a reduction in data rate with little loss in perceived quality.

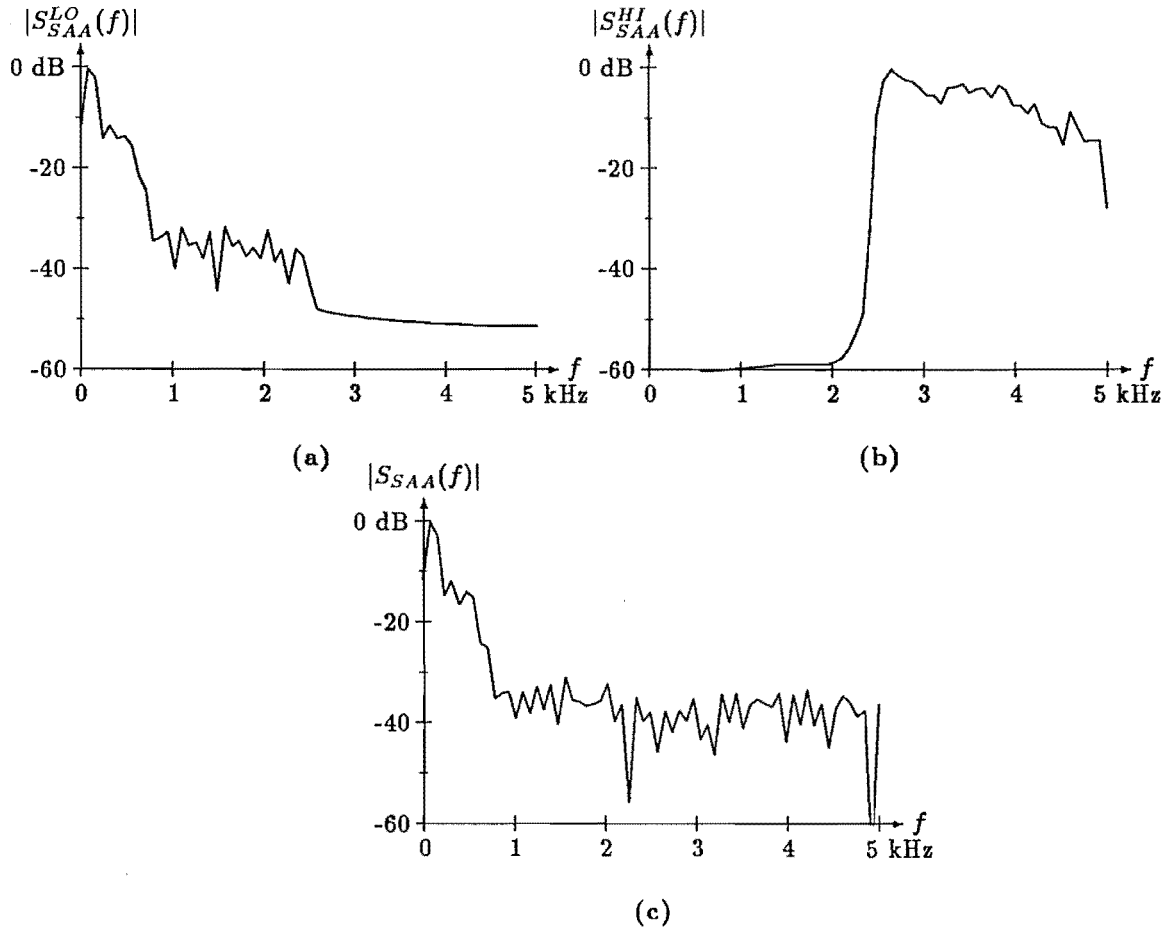


Figure 5.38. Specta of SAA signals obtained from low and high frequency sub-bands of the utterance AM-RAIN1. **a:** 0–2.5kHz, **b:** 2.5–5kHz. Note that the filters employed to separate the sub-bands have a stop-band rejection of 57dB, which is why the spectra shown here have a small component present in the filtered-out sub-band. **c:** Spectra of SAA signal computed from the entire utterance AM-RAIN1. Note that the vertical scales on these spectra are necessarily arbitrary, because of the normalising effect of SAA.

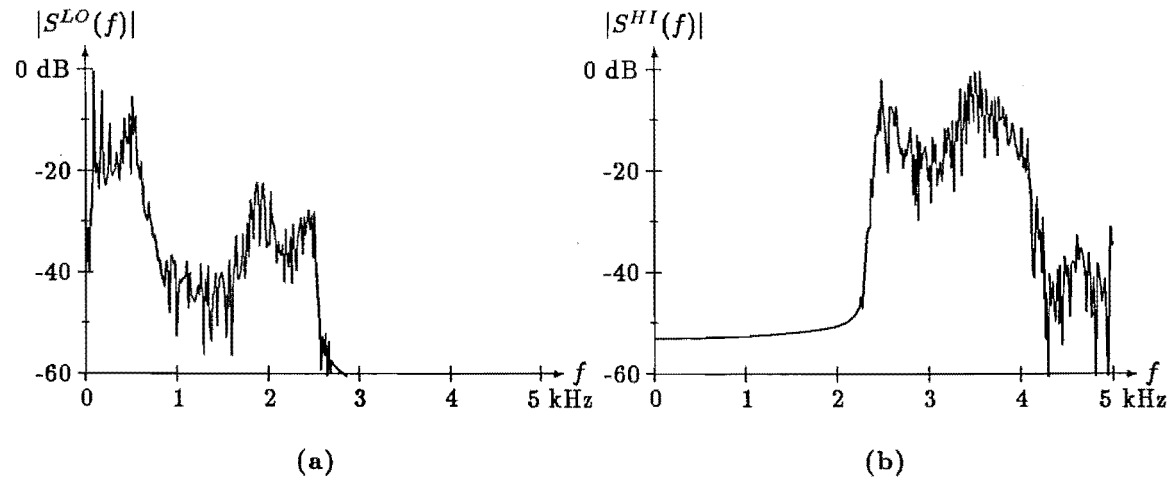


Figure 5.39. Spectra of **a:** low frequency and **b:** high frequency sub-bands of the segment of speech shown in Fig.5.33.

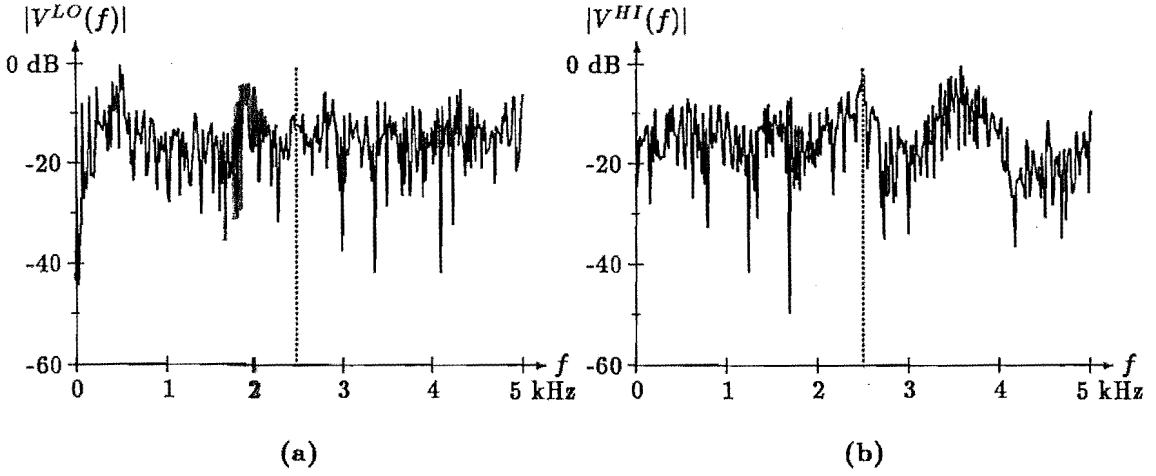


Figure 5.40. Spectra of CLEAN signals obtained from the **a:** low frequency and **b:** high frequency sub-bands shown in Fig.5.39. The number of CLEAN pulses is 1300 in each case, with SNRs of 11 and 10dB respectively. The vertical line in each spectrum indicates the boundary frequency between the sub-bands.

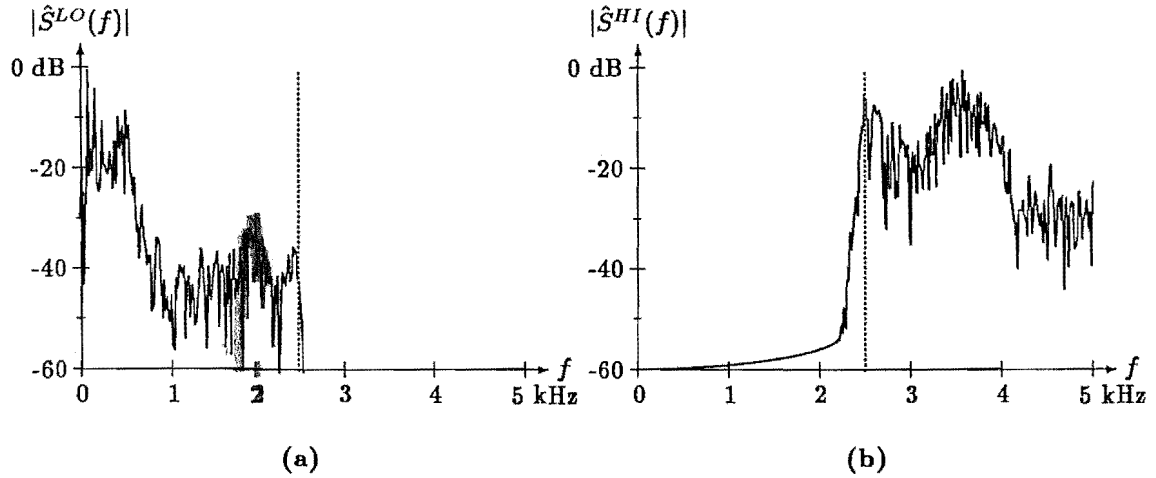


Figure 5.41. Spectra of reconstructed signals formed from the CLEAN signals represented in Fig.5.40.

5.4.3.4 Sub-sampling with "matched" reconstruction filter

One further way to view CLEAN is as a sub-sampling process. Because short-term speech spectra are relatively narrow-band in comparison with the sampled signal bandwidth, there is much redundancy between closely spaced samples of the speech waveform. Many low data rate speech waveform coding schemes (e.g. predictive coding, §3.5.1.1 or sub-band coding, §3.5.1.2) attempt to reduce the data rate by taking advantage of this redundancy.

SAA/CLEAN can be interpreted as another method that takes advantage of the redundancy in speech signals. As discussed in §5.4.3.1, performing SAA and CLEAN "flattens" the spectrum of a speech signal by decomposing it into a finite sum of weighted sinusoids in the frequency domain. These sinusoids correspond to discrete pulses in the time domain. CLEAN can therefore be thought of as a technique to choose a (small) set of pulses (or samples) that, when filtered again by the CLEAN

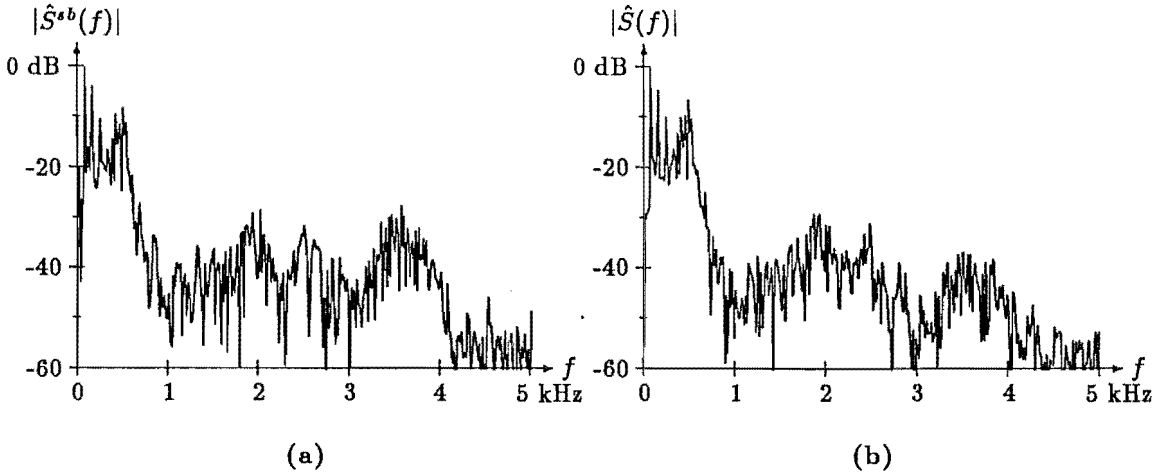


Figure 5.42. Spectra of reconstructed signals formed from: a: The CLEAN signal depicted in Figs.5.34c and 5.35c. b: The superposition of the reconstructions of the two sub-bands.

kernel, "fit" the speech spectrum. In other words, deconvolving the SAA signal from the speech signal by means of the CLEAN algorithm tends to concentrate the energy in the speech signal into a few discrete pulses, which can then be treated as a sequence of (non-uniformly spaced) samples. The spectral "interpolation" which CLEAN extends over parts of the speech spectrum which have little energy can be thought of as an analogue of the spectral folding that occurs in uniform sampling (§1.2.5.5). The aliased energy at these frequencies is suppressed when the CLEAN signal is convolved with the SAA signal to form the reconstructed signal (Fig.5.42).

Viewed as a sub-sampling scheme, SAA/CLEAN can be compared to some other techniques of representing a signal by non-uniformly spaced samples. An early speech encoding scheme of this type is described by Mathews (1959), in which only the extrema of the speech waveform are encoded. Synthetic speech is subsequently generated by interpolating between these points with cubic spline functions. Another description of non-uniform sampling techniques is presented by Yao and Thomas (1967), who discuss the use of Lagrange interpolation functions. The type of non-uniform sampling described in these two papers differs from SAA/CLEAN in that the samples actually represent the amplitude of the signal at the sampling instants. Interpolation functions that are zero at all of the other sampling instants must therefore be employed. The CLEAN pulses are not restricted in this way, so the correct signal amplitude at the "sampling instants", as well as at all other instants, must be reconstructed by convolving the CLEAN pulses with the CLEAN kernel. This difference means that the CLEAN signal cannot be called a (non-uniformly) sampled signal in the traditional meaning of the term "sampled signal".

Chapter 6

Asthmatic cough analysis system

This chapter presents the results of my research into methods of analysing asthmatic cough sounds. The purpose of the research presented here is to design and implement a micro-computer signal analysis system to facilitate a clinical investigation of the sounds of coughs from asthmatic and non-asthmatic children. The motivations for this type of investigation are introduced in §6.1. The emphasis in this chapter is on the practical details of implementing the analysis system. Some preliminary results are presented in §6.2.1. However, the "wider" research effort, of which this system is an integral part, is still underway. The ways in which it is planned to use this system in the context of the wider research effort are outlined in Chapter 8.

The goal of the wider research is to investigate the characteristics of the cough sounds made by asthmatic and non-asthmatic children. It is hoped that these characteristics can give an indication of the presence and severity of the asthmatic condition, particularly in young children. This would be clinically useful because other methods of assessing the severity of asthma are often impractical for this group of patients (§6.1.1).

§6.1.3 describes the physiological mechanisms of cough and cough sound production. §6.2 then discusses the methods of analysis employed to reveal the characteristics of the sounds. The cough analysis system, introduced in §6.3 and §6.4, contains all the necessary hardware and software to collect cough sounds, perform spectrographic analysis on segments of the sounds and examine the time domain and spectrographic signals in detail. In

6.1 Background

6.1.1 Introduction to asthma diagnosis

The diagnosis of asthma is usually made on a historical basis and by considering the clinical symptoms exhibited by the patient (cf. Connolly and Godfrey, 1970; Cloutier, 1983; Canny and Levison, 1987). In addition, there are several methods that are commonly employed by physicians to assess the degree to which the air passages are affected in asthma. These are usually based on a *provocation* or *challenge test*, where an asthma-inducing stimulus is given to the patient and the airway response is evaluated (cf. Josephs *et al.*, 1990). Conventional stimuli include exercise (e.g. several minutes of free running, cf. McFadden, 1984; Tsanakas *et al.*, 1988) or the inhalation of an irritant substance (e.g. Histamine or Methacholine, Jones, 1966). The airway response is usually evaluated indirectly by means of so-called "spirometry", which measures the change in airways resistance. This is done by asking the patient to forcefully exhale. The differences in peak airflow and/or exhaled volume between the "pre-challenge" and "post-challenge" test characterise the change in airways resistance, which in turn

indicates the severity of the asthmatic condition (Jones, 1966; Josephs *et al.*, 1990).

Spirometric measurements commonly made include the *peak expiratory flow rate* (PEFR), which is the peak flow reached during the forced expiration, the *forced expiratory volume* (FEV), which is the total volume of air exhaled, and the *FEV1*, which is the volume exhaled during the first second of an expiration (cf. Cotes, 1975, Chapter 5). PEFR and FEV1 are significantly reduced in patients with asthma during and sometimes even between symptomatic episodes of wheezing (cf. Jones, 1966).

Challenge test methods provide a quantitative and reproducible indication of the presence and severity of asthma in older patients. However, for young children and infants, several problems arise which often render this approach unsuitable. Firstly, it is often difficult to persuade children to co-operate and perform the necessary exercise. Secondly, it is usual for the evaluation of lung function to require a maximal forced expiration (Tsanakas *et al.*, 1988). With young children (particularly those of less than four years) it is often difficult to ensure that the expiration is produced with maximal effort. Avital *et al.* (1988) suggest that the presence of wheezing can be used as a reliable indicator of increased airway resistance in young children, for whom conventional spirometry is inappropriate. A third factor is that young children with mild or developing asthma may only exhibit symptoms, such as the occurrence of persistent or night-time cough, which cannot be readily observed in the clinic (cf. Corrao *et al.*, 1979; Cloutier, 1983; Anonomous, 1988).

6.1.2 Other relevant research

The occurrence and nature of coughs associated with asthma attacks in children, especially those occurring at night, have been the subject of recent studies (Archer and Simpson, 1985; Toop *et al.*, 1986). The research presented here is a continuation of these studies, with signal processing techniques being introduced to allow more sophisticated analyses of the cough sounds to be made (Toop *et al.*, 1989a). Previous studies of asthmatic cough sounds in children have compared the numbers of coughs occurring during the night with the severity of the asthmatic condition (Archer and Simpson, 1985; Toop *et al.*, 1986; Thomson *et al.*, 1987).

The characteristics of cough sounds in general have also received some attention. Kelemen *et al.* (1987) compare plots of air flow rate, lung volume, and sound, all versus time, for different types of cough. They, and Korpas *et al.* (1987), describe cough sounds according to their waveforms, finding that the envelopes of the sounds appear to differ between patients with different diseases. Further details of their cough classification scheme are summarised in §6.1.4.3. Korpas *et al.* (1987) and Salát *et al.* (1987) characterise a cough sound by its energy, which is the integral of the sound intensity. They present results which indicate that there is a significant difference between the total energy of asthmatic and non-asthmatic coughs.

The spectral content of cough sounds has only recently received attention (Debreczeni *et al.*, 1987,1990; Toop *et al.*, 1989a; Piirilä and Sovijärvi, 1989). Both Debreczeni *et al.* (1987) and Piirilä and Sovijärvi (1989) obtain the average spectrum of the cough sound, while Toop *et al.* (1989a) and Piirilä and Sovijärvi (1989) compute spectrograms to indicate the evolution of spectral components with time. Debreczeni *et al.* (1987,1990) compare the coughs of groups of patients with different types of airway disease. They treat each spectral component separately, using a statistical test to determine which frequency bands contain significantly different amounts of energy between each pair of groups. Although their average asthmatic and non-asthmatic cough spectra appear different, the wide inter-patient variability means that only a few frequency components are significantly different.

Piirilä and Sovijärvi (1989) estimate the upper frequency limit for the average

spectrum of each cough. In addition, they estimate the duration of the first and any subsequent sound, and the occurrence and duration of any "wheezing components", which they define as 'continuous sounds with duration more than 250 ms'. They find that the upper frequency limit is lowest for asthmatic coughs, and that the wheeze duration, as a proportion of the total cough duration, is greater in asthmatic than in other types of cough.

Several studies have investigated the spectral content of lung and breath sounds for asthmatic and other conditions (Akasaka *et al.*, 1975; Gavriely *et al.*, 1981; Cohen and Landsberg, 1984; Baughman and Loudon, 1985; Fenton *et al.*, 1985; Pasterkamp *et al.*, 1989). These have been attempts to quantify the lung sounds that physicians have for many years employed as an indication of the presence of asthma and various other ailments (Forgacs, 1978; Loudon and Murphy, 1984). Aspects of these investigations are introduced whenever relevant throughout this introductory section.

6.1.3 Coughs

The physiological purpose of coughing is to expel foreign particles and excess mucus from the airways (cf. Cloutier, 1983). Coughing is thus necessary for survival and occurs in all people, healthy or not. However, in some cases, pathologies of the airways cause *pathological* or *chronic* coughing (Irwin *et al.*, 1977; Cloutier, 1983). The common cold and, as has already been suggested, asthma are two well known examples of conditions which can cause chronic cough (Cloutier, 1983).

Non-voluntary coughs occur when the *cough reflex* is triggered (Macklem, 1974). This usually happens when some sort of irritation occurs to a nerve ending of one of the many *cough receptors*, which are spread throughout the airways (Irwin *et al.*, 1977). The act of coughing usually consists of the following physiological actions. Firstly, air is inhaled into the lungs. Secondly, the glottis closes and the breathing muscles compress the lungs, producing a high air pressure in the sub-laryngeal airways. The glottis then opens suddenly, and air is forcefully expelled through the mouth (Irwin *et al.*, 1977).

The compression of the *pleural cavity* (the space between the rib cage and the lungs) exerts a compressive force on the lungs and airways. Because these have an elastic recoil, this force increases the pressure in the airways to be above atmospheric pressure. Hence the air is forced out of the lungs and the airways (Macklem, 1974). Fig.6.1*a* depicts the lungs and airways diagrammatically during a cough. The pressure in the lung P_{alv} is greater than the pleural pressure P_{pl} because of the elastic recoil of the lung. The pressure at the mouth P_{atm} is much less than P_{pl}, and so the pressure difference across the walls of the airways varies as shown in Fig.6.1*b*. Because of the compressive forces on the airways downstream of the equal-pressure point (where $P_{pl} = P_{alv}$), the airways tend to collapse, thus increasing their resistance and causing the steep drop in pressure shown in Fig.6.1*b* (Macklem, 1974). In some diseases such as asthma, the airways are already partially occluded, and so the equal-pressure point occurs more distally (further from the mouth). Hence the airways are usually compressed to a greater extent in such conditions.

The compression of the airways implies that there is a limit on the maximum airflow that the airways can support. This limitation occurs because increasing pleural pressure (necessary for higher flows) is counteracted by greater airways resistance caused by the collapsing airway walls (Dawson and Elliott, 1977). The velocity of the air flow through the airways is increased dramatically because of this compression of the airways. Ross *et al.* (1955) measure the compression by means of x-ray tracings of the airways. They estimate that the diameter of the larger airways can be compressed by as much as 50% during a cough by a healthy subject. Air velocities of up to 120m/s have been measured within the airways during a cough (Macklem, 1974). However,

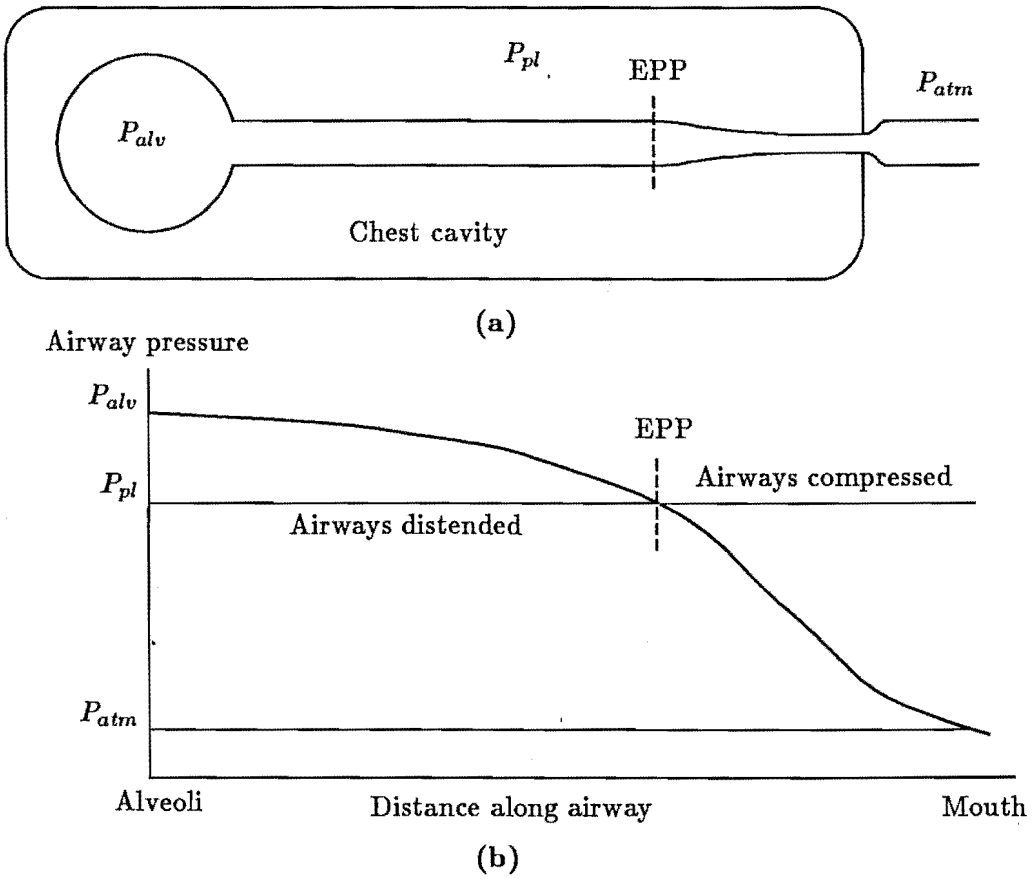


Figure 6.1. **a:** Model of lungs and airways during a cough. The pressure in the lung is greater than the pleural pressure because of the elastic recoil of the lung. **b:** Diagram showing the pressure gradient across the airway walls at different positions. The airways become compressed downstream of the equal-pressure point

even though such high velocities may be achieved, the volume flow of air is limited because of the narrow cross-section of the airways (cf. Macklem, 1974; Dawson and Elliott, 1977).

Asthma is characterised by swelling of the mucosal lining of the airways. This increases the thickness of the airway walls, and decreases their internal cross-sectional area. Hence the resistance of the airways is increased, and the peak air flow is reduced (cf. Loudon and Shaw, 1967). Excessive mucus production also occurs and this, together with the swollen mucosa, may mechanically stimulate the cough receptors (see the first paragraph of this section).

6.1.4 Cough sounds

This section reviews the sounds that accompany coughing. In §6.1.4.1 I introduce the models invoked to explain the mechanisms by which sounds are generated during coughs, while in §6.1.4.2 I briefly discuss the way in which the transmission of sounds through the lungs and airways is affected by pathological conditions such as asthma. §6.1.4.3 outlines the subjective terms given to cough sounds by physicians, and then introduces the schemes that have been proposed to quantitatively describe various types of cough sounds.

6.1.4.1 Models of cough sound production mechanisms

Sound energy during a cough is generated both by turbulence in the air flow, and by "flapping" of the airway walls. The literature on models that describe these sources of sounds is sparse, although models of airflow during coughing or breathing (§6.1.3) form the basis of descriptions of sound generation (cf. Forgacs, 1978; Grotberg and Davis, 1980; Gavriely *et al.*, 1984; Webster *et al.*, 1985).

In this section I briefly introduce one approach to understanding how cough sounds are generated. The purpose of this discussion is to provide some justification for the types of analysis described in §6.2. Hence I provide only a broad outline of the important concepts. The mathematical details are presented in greater depth by Grotberg and Davis (1980) while discussion of the interpretaion of breath sounds in light of this model of sound production is provided by Gavriely *et al.* (1984).

The mechanisms by which cough sounds are generated can be modelled by considering the airways as flexible channels and by analysing the flow of air through such channels (Grotberg and Davis, 1980). Such an analysis can be applied to the flow of air arising from both breathing and coughing. It also enables the effects of changes in the airway structure on the airflow (such as those that occur in asthma) to be investigated. By means of these analyses, the types of sounds produced under different conditions can be predicted (Gavriely *et al.*, 1984; Hinchey and Snellen, 1987).

Because the airway walls are flexible, they form an unstable mechanical-aerodynamic system when the air velocity is greater than some critical value (Grotberg and Davis, 1980). This instability can be simplistically understood as arising from the interaction between the forces generated by the movement of the air, the pleural compression of the airways, and the elastance and mass of the airway walls. As the airways are compressed (flattened) by the pleural compression, the air velocity increases as described in §6.1.3. In addition, the Bernoulli force which the movement of the air induces also causes the airways to flatten. At some critical velocity, the airway walls begin to vibrate ("flap"). The vibration occurs because of an unstable relationship between air velocity and airway cross-section: Increasing air velocity causes the airway cross-section to decrease, but this increases the resistance of the airways, thus reducing the air velocity. This flapping of a flexible tube can be easily observed by blowing up a balloon and then letting the air escape. The neck of the balloon (the flexible tube) vibrates according to the above mechanism. The frequency of the flapping depends on the elastance, mass, and size of the airway walls (Forgacs, 1978; Grotberg and Davis, 1980; Gavriely *et al.*, 1984).

The flapping of the walls occurs together with a vibration in the airflow through the airway, which vibration constitutes the "wheeze" type of sound often heard when the airways are constricted by, for instance, asthma. Asthma causes the airway walls to thicken, thus reducing the airway size, increasing the wall mass, decreasing its elastance, and increasing the air velocity. Hence the critical velocity is reduced and vibrationary sounds are more easily generated (Gavriely *et al.*, 1984).

The model described above of flow through a flexible tube also explains the airflow limitation mentioned in §6.1.3. This occurs because the increase in air velocity is matched by a consequent decrease in airway cross-section (cf. Dawson and Elliott, 1977). Hinchey and Snellen (1987) suggest that airway vibration is almost always associated with flow limitation. They, and Beck and Gavriely (1990), find that the onset of flow limitation is often accompanied by the appearance of high frequency (1–2kHz) peaks in the short-term spectra of the breath sounds.

In addition to the "oscillatory" types of sounds generated by the flapping of the airway walls, "noise-like" sounds are generated by turbulence in the airflow. Turbulent flow occurs when the air velocity exceeds a critical value, that is dependent on the

Reynolds number of the gas. Turbulence can be precipitated by bifurcations or obstacles in the airways, or by surface "roughness' on the airway walls (Reynolds, 1974, Chapter 1). The airway narrowing which occurs during asthma because of the swelling of the airway walls increases the air velocity, which leads to greater air turbulence. Asthma is also associated with the excretion of mucus, which changes the surface characteristics of the airway walls. Hence the character of the air turbulence is likely to be different than in the non-asthmatic case. For instance, asthmatics commonly exhibit a type of cough which is termed "moist cough", referring to its association with mucus secretions.

6.1.4.2 Sound transmission through the lungs and airways

The character of the sounds are modified by their transmission through the body before they can be recorded. There are two routes for the sounds to travel. A microphone placed on the chest wall picks up sounds that propagate through the lung tissue and chest wall. Wodicka *et al.* (1989) model the transmission characteristics of such tissue by considering it to be constructed of many small bubbles of air, distributed throughout a medium that essentially has the acoustic properties of water. Because of the consequent absorption of acoustic energy by such a structure (Wodicka *et al.*, 1989), the sounds that pass through the tissue are strongly attenuated, especially those of higher frequencies. The sounds heard on the chest wall therefore consist almost entirely of the low frequency components of the original sounds. Forgacs (1978, p11) shows that the healthy lung acts like a low pass filter with a cut-off frequency of about 200Hz and an attenuation that increases at about 10–20dB/octave above that frequency.

Sounds that are detected at the mouth must propagate through the airways and mouth. This is essentially the vocal tract described in §2.3.1.3, although with the important addition of the sub-laryngeal airways. The influence of the supra-laryngeal vocal tract on the cough sounds is not likely to change much between coughs because the mouth is usually relatively open during a cough. However, the sound transmission characteristics of the sub-laryngeal airways are likely to be affected by asthma because of the narrower, thicker airways and the increased amount of mucus on the airway walls (Gavriely *et al.*, 1984).

6.1.4.3 Types of cough sounds

Since everybody coughs, and for many different reasons, the value of a cough as a diagnostic aid depends upon the ability of the physician to distinguish between *normal* and *pathological* coughs. The frequency and persistence of coughing is one indication that the cough is caused by some pathology (Irwin *et al.*, 1990).

The characteristics of the sounds of different types of cough have been described in subjective terms by physicians as an aid to categorising their causes. Some of these subjective terms are metaphorical, such as the terms "wheezy", "tight", "brassy", "squeaky", and "rasping", to name a few (Toop, 1989). Other coughs are simply labelled by association with a particular disease. For example, a physician may describe a cough as "asthmatic" or "barking" (in the case of croup), thus indicating a characteristic type of cough (Toop, 1989). The particular characteristics of each of these terms are of course learnt by experience, although the metaphorical terms provide a basis for categorising the different types of cough.

The metaphorical terms mentioned above for the different cough sounds indicate that the coughs arising from different diseases do indeed "sound different" to the ear. It is likely that the differences between coughs with names such as "wheezy" and "brassy" are largely related to differences in the spectral content of the sounds. There-

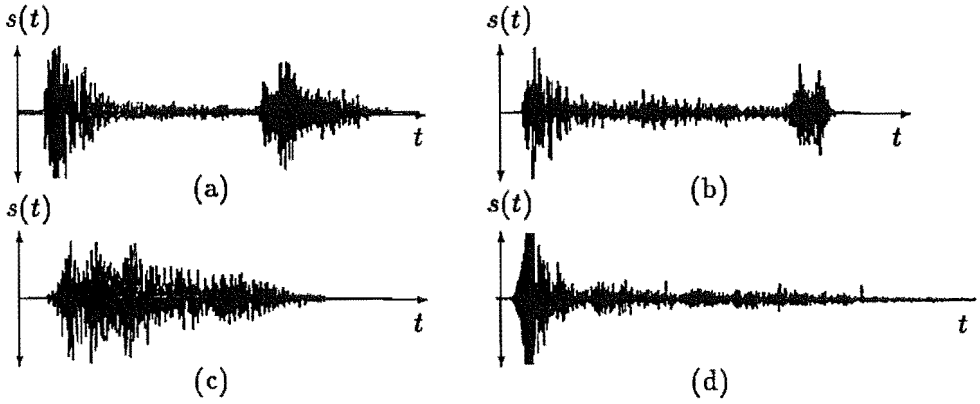


Figure 6.2. Waveforms of typical cough sounds. The coughs shown in in **a:** and **b:** exhibit the three phases referred to in the text as the "first sound", the "noisy interval", and the "second sound". **c:** Cough sound with an extended initial burst. **d:** Cough sound having no final burst.

fore, spectral analysis is a logical approach to obtaining a quantitative description of the different "types" of cough.

One of the difficulties in designing a quantitative analysis procedure for describing the cough sounds is the translation of the subjective terms into equivalent acoustic features. Kelemen *et al.* (1987) divide a cough sound into seven parts, with three major phases. These three phases are the "first sound", the "noisy interval", and the "second sound". They relate the first sound to the rapid acceleration of flow at the start of the cough, the noisy interval to the relatively steady flow during the cough, and the final sound to the flow deceleration as the cough ends. Korpas *et al.* (1987) suggest that the final sound originates in the larynx. In any particular cough, the second and/or third phase may be missing. Fig.6.2 shows typical cough waveforms that exhibit the three phases.

Korpas *et al.* (1987) show representative waveforms of cough sounds corresponding to different diseases. They describe cough sounds as generally consisting of a "first" and a "second" cough sound. This pattern changes for different diseases, so that a particular sound may be composed of one to four or more "bursts", with the "gaps" between each burst also varying according to the disease.

6.2 Analysis Methods

The aim of analysing cough sounds is to classify the coughs according to physiological changes in the airways. Developing signal processing techniques for analysing the cough sounds involves three major steps. Firstly, the types of features which characterise the coughs and their differences need to be established. Secondly, techniques for extracting and quantifying the features must be developed and refined. The third step is to develop a system by which the severity of any clinical condition can be automatically inferred from the features of the cough sound.

§6.2.1 describes the preliminary analysis (Toop *et al.*, 1989a) of several examples of asthmatic and non-asthmatic cough sounds which was carried out in order to justify the development of the cough analysis system presented in §6.3 and §6.4. In §6.2.2 I explain why spectral analysis techniques are invoked to characterise the cough sounds. The several types of descriptive features that can be used to characterise the spectra of the sounds are discussed in §6.2.3. The third step mentioned above, of automatically inferring the severity of asthma from the features of the sounds, is not mentioned further in this chapter and is relegated to §8.2.3.

Label	Age	Asthmatic?	Time recorded	PEFR reduction
A-PRE	7	Yes	Pre-exercise	normal
A-POST	7	Yes	10 min. post-exercise	30% fall
N-PRE	7	No	Pre-exercise	normal
N-POST	7	No	10 min. post-exercise	10% fall

Table 6.1. Coughs employed in the preliminary analysis of cough sounds. Coughs from two children are presented here. The labels identify the child ("A" or "N") and the time at which the sound was recorded ("PRE" or "POST" exercise). The PEFR reduction column refers to the measured PEFR as compared with the expected normal value for a child of that height.

6.2.1 Preliminary analysis of cough sounds

As mentioned in §6.1.2, several researchers have examined the spectral content of lung sounds. In order to determine if spectral analysis of the cough sounds might be a useful avenue to take in an investigation of their characteristics, I recorded and analysed several examples of asthmatic and non-asthmatic coughs. §6.2.1.1 describes the methods employed in this analysis, while §6.2.1.2 presents the results so obtained.

6.2.1.1 Methods invoked in the preliminary analysis

Coughs were recorded from a seven year old boy with a history of asthma and nocturnal coughing and also from a child with no asthmatic history. Coughs were recorded from both children before and at intervals after undertaking strenuous exercise. The cough sounds were recorded using a SONY ECM-16T condenser lapel microphone connected to a UHER 4200 tape recorder. From this record, selected examples of coughs were digitised at a rate of 20kHz by a 12bit LPA11-K A/D converter, and stored in a DIGITAL VAX computer. The signals were low-pass filtered by a KEMO VBF/8 filter, having a 3dB cutoff frequency of 9kHz and a roll-off of 48dB/octave above that. Table 6.1 identifies the coughs involved in this experiment, together with the condition of the patient at the time each one was recorded.

Subsequent to digitising the coughs, it was found that their spectral components above 5kHz contained negligible energy. Each one was therefore digitally filtered, to remove all energy above 5kHz, and re-sampled at a rate of 10kHz. This re-sampling was done to reduce the disk storage requirements.

The time-varying spectrum of each cough sound was computed by the method described in §3.3.1. Segment of 25.6ms in duration, each one starting 5ms after the start of the one before, were multiplied by a 3-term Blackman-Harris window and were then Fourier transformed. The resulting spectrogram, which comprises all the short-term spectra computed as described above, is displayed as a "stack plot" in order to reveal the evolution of spectral components with time. A stack plot produces a "three-dimensional" effect, with frequency as the horizontal axis, time "receding" up the page as if into the distance, and the spectral magnitude at each epoch represented by the vertical height of the short-term spectrum at that epoch above an imaginary base-line.

6.2.1.2 Results of preliminary analysis

Fig.6.3*a* shows the sound waveform and spectrogram of a typical cough (N-PRE) from the normal child before exercise. There is power across the range of frequencies from 500 to 2000Hz with few definite peaks. Fig.6.3*b* shows that, following exercise, the sound

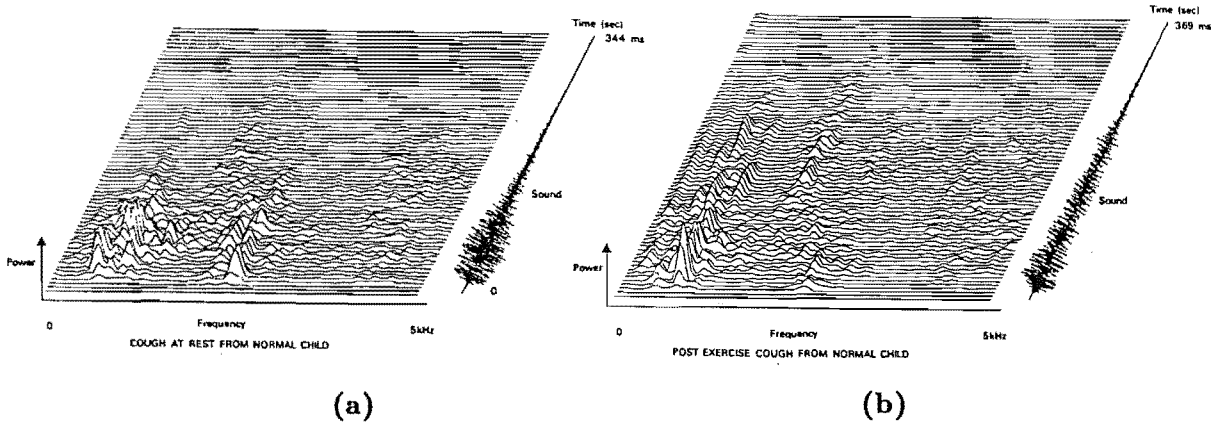


Figure 6.3. Cough sound and spectrogram from normal child. **a:** Pre-exercise cough N-PRE. **b:** Post exercise cough N-POST. Table 6.1 identifies each of the coughs.

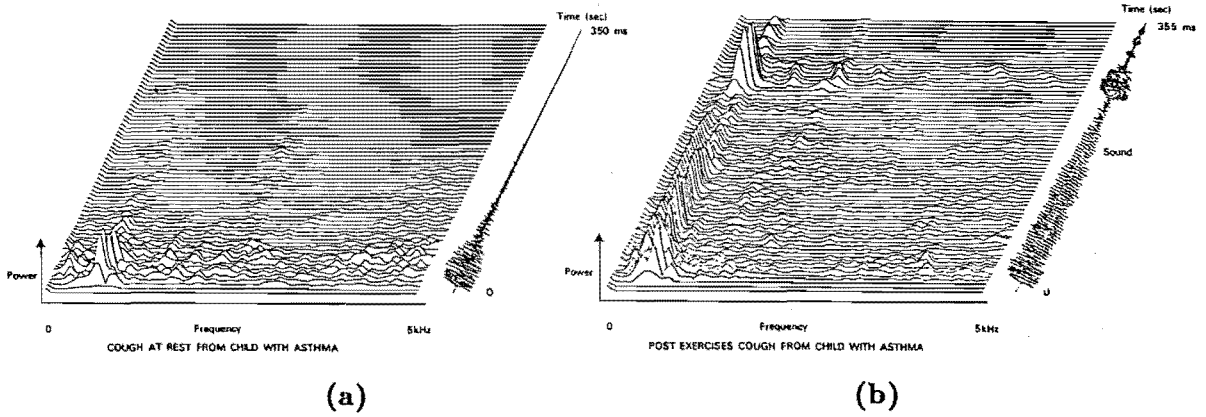


Figure 6.4. Cough sound and spectrogram from asthmatic child. **a:** Pre-exercise cough A-PRE. **b:** Post exercise cough A-POST. Table 6.1 identifies each of the coughs.

of the cough changes little, although the duration of the cough increases somewhat, with the spectral energy also being more "spread out" rather than being concentrated in the initial "burst".

By contrast, the sound of the cough from the child with asthma exhibits major changes following a similar interval of exercise. The sound and the spectrogram for the pre-exercise cough A-PRE, shown in Fig.6.4*a*, are similar to those of the pre-exercise cough shown in Fig.6.3*a*. The sound is of short duration and the spectrogram indicates energy in a wide band of frequencies. However, the post-exercise cough A-POST, shown in Fig.6.4*b*, is very different from its pre-exercise counterpart. The duration of the cough increases significantly, and the spectrogram indicates much sharper concentration of energy in spectral components around 500–600Hz. In addition, there is a second "burst" of sound some 250ms into the cough. This sound has a very pronounced peak at a frequency of about 600Hz, together with what appear to be harmonics at 1200 and 1800Hz.

6.2.2 Discussion of the merits of spectrographic analysis

During preliminary analyses of several coughs (§6.2.1), features based on the spectral content of the sounds were chosen as most the promising for describing the changes

associated with asthma. Spectral analysis of the sounds enables the occurrence of *wheezes* to be detected. As described in §6.1.4.1, wheezes are produced by vibration of the airway walls. During asthma, the airway walls are thickened, leading to an increase in air velocity and hence a greater likelihood of wheezes occurring. In addition, the frequency of the wheezes depends on the mass and elastance of the airway walls, both of which are changed by an asthmatic condition.

In addition to wheezes, cough sounds contain a wide spectral range of "noise" components generated by turbulence in the airways. One expects that the characteristics of the turbulent noise should alter if the airways change due to asthma. Other "noise" components of cough sounds are generated at the glottis, as it opens and closes at the start and end of each cough. These sounds are likely to have some semblance of a harmonic nature, due to the quasi-periodic glottal vibration. However, they tend to dominate when present since they are generally of much greater intensity than the sounds generated in the airways (see §6.1.4.3 and §6.2.3).

The sound transmission characteristics of the airways (§6.1.4.2) are dependent on the cross-sections of the airways and the nature of the airway walls (cf. §2.3.1.3). Since asthma affects both of these aspects of the airways, one would expect that the sound transmission characteristics would also change during asthma. Specifically, reduction in airway cross-section should lead to an upward shift in the resonant frequencies of the airways.

The purpose of time-varying spectral or *spectrographic* (§3.3.1) analysis of cough sounds is to display the evolution of the spectral content of the sounds. Spectrographic analysis identifies short-term spectral components of sounds during the course of coughs. One difficulty encountered in such an analysis is the unavoidable trade-off between temporal and spectral resolution (§3.3.1). If one wishes to observe the rapid changes that occur during the course of a cough, one must accept a limited resolution of spectral details. This could be a drawback in the identification of "wheezes", which are characteristically narrow-band "tones".

6.2.3 Characteristic features of cough sounds

Although a full description of characteristic features which describe the cough sounds is beyond the scope of this thesis, it is important for the reader to appreciate the general characteristics of the sounds upon which a future system for determining the presence and severity of asthma can be based. In this section I outline the types of characteristic features which I think may usefully describe the differences between cough sounds.

Cough sounds are inherently high intensity, "noisy", sounds, which poses difficulties when one attempts to extract useful information from them. This may by one reason why there has been so few previous analyses of them. However, both from examining the previous literature on the subject (§6.1.4.3), and from the preliminary results of the spectrographic analysis presented in §6.2.1.2, it seems that most of the "noisy" sound energy in the cough appears in the initial and final "bursts". During the second, quieter, phase, more subtle sounds, that are perhaps due to the particular airway characteristics, may be more apparent. In addition, the second part of the cough sound appears to be more interesting from the point of view of investigating changes in the airways because the glottis, being fully open, has little influence on the airflow. Any noises heard during this interval are thus likely to originate in the airways themselves. Indeed, experience suggests that wheezes are heard during this phase (Toop, 1989).

The classification of cough sounds by Kelemen *et al.* (1987) into different types according to their energy envelopes provides only a coarse level of discrimination between different types of cough. Despite this, Kelemen *et al.* and Korpas *et al.* (1987) show some generic differences between cough sound waveforms for different diseases. In

view of the comments made in the previous paragraph, however, I think that examining changes in the spectral content of the cough sounds is a more promising approach to identifying the severity of any changes in the airways.

Chabonneau *et al.* (1983) characterise the spectra of breath sounds in terms of several parameters computed from the average spectrum of each breath. These parameters are the the peak frequency, the half-power bandwidth, the highest significant frequency (the highest frequency with amplitude greater than 10% of the spectral amplitude at the peak frequency), and the weighted mean frequency. An index composed of the sum of normalised parameter values appears to discriminate successfully between asthmatic and normal breath sounds.

Because of gross differences in intensity between the three "phases" of a typical cough sound, Chabonneau *et. al's* (1983) approach to analysing breaths must be modified before it can be adapted to the characterisation of coughs. It is necessary to analyse each phase of a cough sound separately. Debreczeni *et al.* (1987) and Piirilä and Sovijärvi (1989) partially accomplish this by only computing the average spectrum for the part of the cough sound which follows the initial burst. Debreczeni *et al.* (1990) compute average spectra from successive 50ms segments of the cough sounds, finding that later segments are somewhat better at distinguishing coughs from different diseases.

In summary, a system designed to extract characteristic features from cough sounds must, first, separate each cough sound into individual phases and, second, estimate parameters that characterise the average spectra of each phase.

6.3 A clinical cough analysis system

In this section I introduce the system that was designed to facilitate the operation of a clinical study employing the analysis techniques described in §6.2. §6.3.1 gives an overview of the system and the requirements which guided its development. The system hardware is described in §6.3.2 and §6.3.3. The software for the system, which absorbed the bulk of the effort put into constructing the system, is described in detail in §6.4.

6.3.1 System overview and requirements

The cough analysis system is intended to facilitate the operation of the clinical study of cough sounds. To this end, it is designed to be easy to operate, with processing performed automatically wherever possible.

The requirements of the system are, firstly, that it be capable of simultaneously digitising several different signals at possibly different sampling rates. In the clinical trials that have been performed so far with this system, the system has been configured to digitise air flow rate from a flow meter and sound pressure from two microphones. However, when digitising sounds that have been previously recorded on a tape recorder, only a single channel is required. The flow signal is used to align the sounds with the cough, and also to perform standard spirometric tests (§6.1.1) during the recording session. The microphones are placed at the mouth and on the chest wall. Traditionally, physicians have listened to breathing sounds through the chest wall only. However, sounds recorded at the mouth are not affected by the severe high frequency attenuation which occurs through the lungs and chest wall (§6.1.4.2). Also, it is sometimes easier to record sounds at the mouth, such as when night coughs are being obtained (§6.1.2).

The rate at which the system is required to digitise the sounds is 5kHz. This rate was determined by a preliminary study of a few cough sounds (§6.2.1), in which

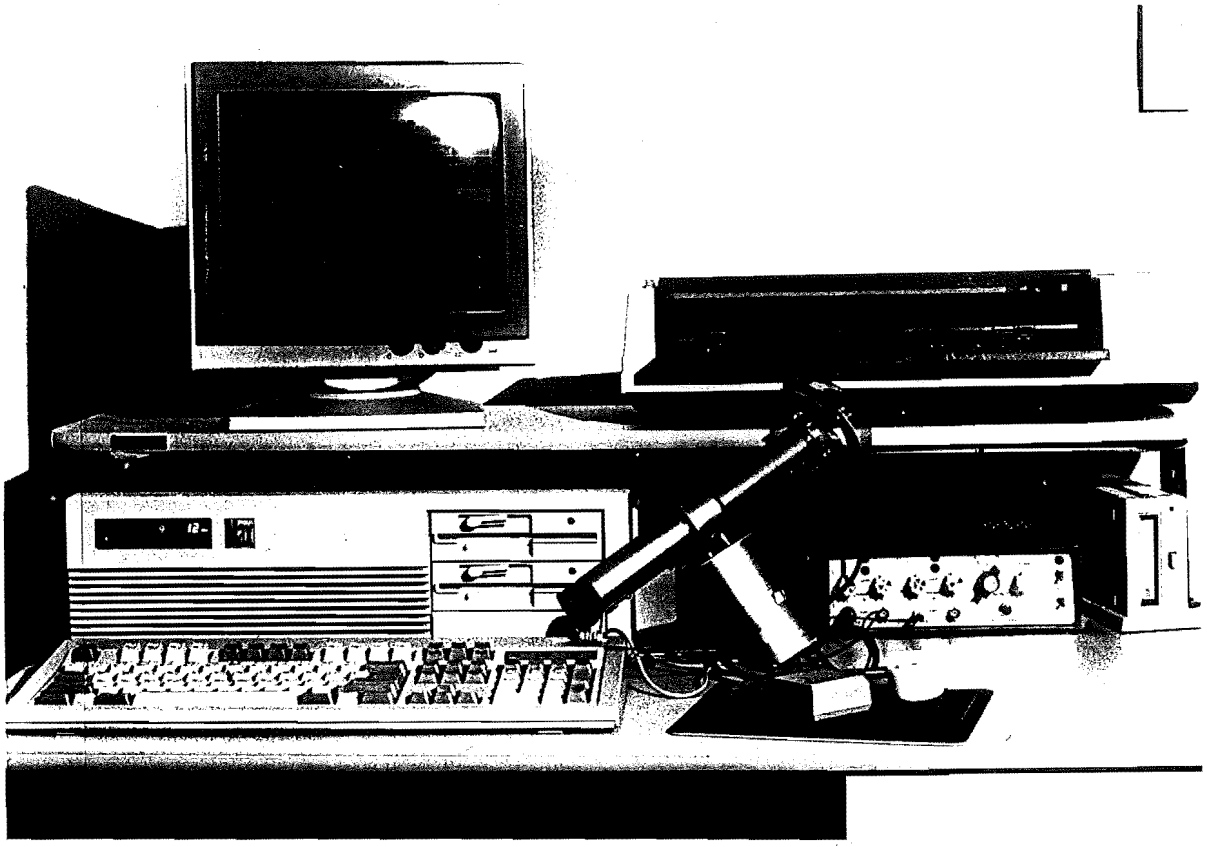


Figure 6.5. Photograph of the equipment comprising the Cough analysis system

it was noticed that most of the energy in the sounds occurs at frequencies less than 2kHz. A much lower sampling rate, of 500Hz, is adequate for the flow channel. This is conveniently one tenth of the sampling rate for the sound channels.

The system is also required to be capable of interactively displaying the signals and allowing the interesting portions to be extracted (from all channels simultaneously) and stored for later analysis. Spectrographic analysis must be performed on the stored sound channels, with the results being displayed on the screen or a plotter. The displays of both the signals and their spectrograms need to be examined in detail.

A third requirement of the system is that it must be as simple as possible to operate. This is particularly necessary in the data collection parts of the system so that the clinicians can concentrate their attentions on the patient rather than the computer. More generally, the system is required to store and process a large amount of data from many patients. In order for the research to proceed efficiently, it is necessary that the data be managed in a fashion that is comprehensible to the user and discourages mistakes.

A final requirement of the system is that it be expandable. Future developments that are planned include the automatic identification of the end-points of individual coughs, the extraction of descriptive features from the spectrographic representations of the cough sounds, and, if the research proves it to be possible, the extraction of an "asthmatic severity index" from the cough sounds (see §8.2.3).

6.3.2 System Hardware

A photograph of the system is shown in Fig.6.5. The flow meter was built by the Department of Bio-Engineering and Medical Physics, Christchurch Hospital. It mea-

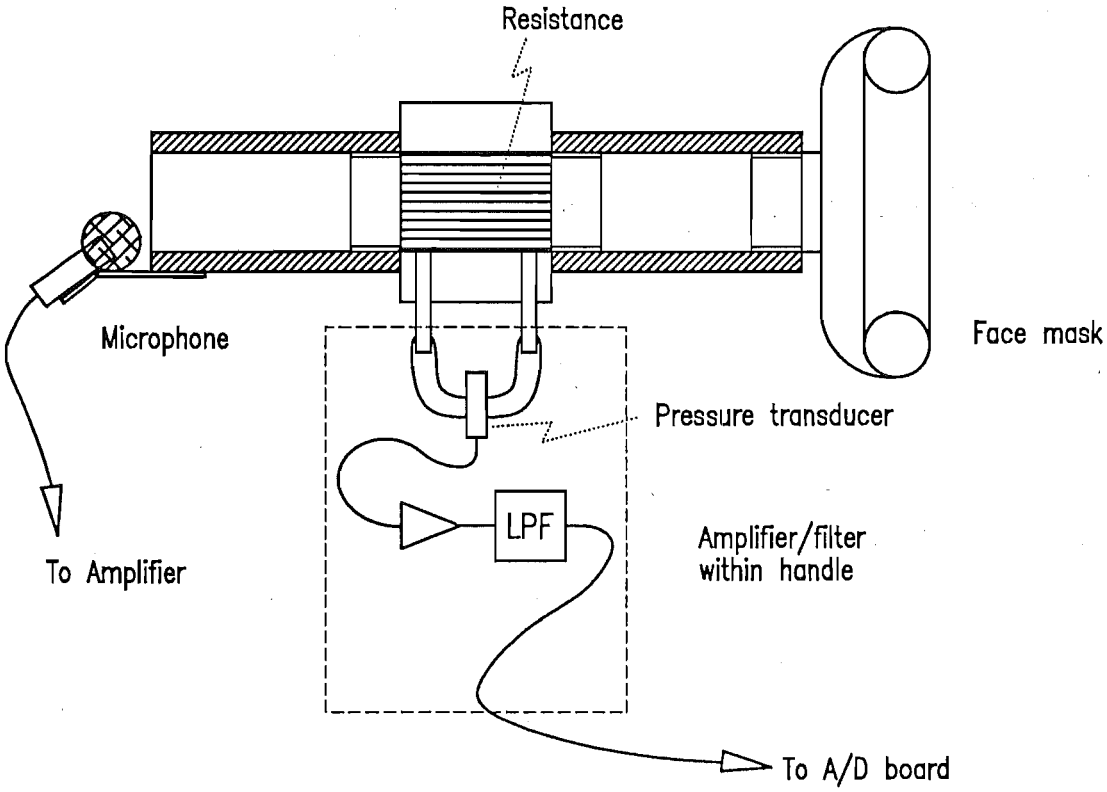


Figure 6.6. Cross-sectional view of the flow meter used to measure the air flow during coughs.

sures airflow indirectly by sensing the pressure differential developed across a small aero-dynamic resistance introduced into the airflow. The resistance is constructed of a spiral coil of corrugated stainless steel plate. The pressure differential is detected by a piezo-electric pressure transducer. The signal from the flow meter is amplified and low-pass filtered to 100Hz. A face mask is used to direct cough air flow into the flow meter. The flow meter was calibrated by connecting it in series with a calibrated flow meter and comparing the A/D sample values to the flow values provided by the standard. §6.3.3.1 describes the calibration procedure.

The chest sounds are obtained through a HEWLETT-PACKARD HP-21050A contact microphone. It has a flat frequency response between 30 and 2000 Hz. The mouth microphone is a BEYER-DYNAMIC MCE-6.9 condenser microphone which has a frequency response that is flat between 20Hz and 10kHz. The mouth microphone is mounted on the exhaust end of the flow meter. Mounting the microphone in this way does mean that the sounds are affected by their passage through the face mask and flow meter. However, all the coughs are affected similarly, so the distortion can be estimated and removed by post-processing within the computer. §6.3.3.2 describes the procedure that was followed to estimate the acoustic transfer function of the flow meter and face mask.

The two microphone signals are amplified and then filtered by custom built (by the Bio-Engineering and Medical Physics Dept. at Christchurch Hospital) 7-th order elliptic switched-capacitor anti-aliasing filters having cutoff frequencies of 2.5kHz and stop-band attenuation of 70dB. The gain of the two amplifiers can be controlled to match the dynamic range of the cough signals to that of the A/D converter. Indicators on the front panel of the amplifier unit indicate when the output signal exceeds half (green) and full scale (red) of the A/D input range. A variable gain is necessary because

Constant	Description	Value
S	Annubar flow coefficient	0.62
D	Pipe internal diameter	1.049 inches
ρ	Air density	0.07649 lb/ft^3 at 20°C
N	Units conversion constant	36.91
K	Overall constant	35.10 l/s/(mm)$^{1/2}$

Table 6.2. Values of the constants in (6.1), taken from the manufacturer's data sheet.

of the wide variation in sound intensity between the coughs of different people. The disadvantage of variable gain is that absolute intensities of different coughs cannot be compared. The amplifiers also have provision for feeding in sounds from another source such as a tape recorder. An anti-aliasing filter, amplifier, and speaker is included so that sounds may be replayed from the computer via a D/A convertor. The interface hardware is an ANALOG DEVICES RTI-815-A with 16 12 bit A/D convertors and two 12 bit D/A convertors.

The micro-computer system consists of a SUNDOX 286 computer, which is fully compatible with an IBM PC-AT, running at a clock speed of 12MHz. A HERCULES graphics display and GENIUS GM-6 "mouse" is employed for interaction with the human operator (see §6.4.3). A NEC 114MByte hard disk provides storage for data and software, and an ARCHIVE XLE 40MByte tape streamer is employed for data backup and archiving. A FACIT 4551 plotter is used to produce plots of the processed signals. The entire system fits on a wheeled trolley so that it can be easily moved between different locations in the hospital.

6.3.3 Hardware calibration procedure

6.3.3.1 Calibration of flow meter

The flow meter was calibrated by connecting it in series with an ANNUBAR flow meter (type 713-316ss, size 1.049) and passing a range of constant air flows through both instruments. The pressure differential h_w (in mm of H_2O) through the ANNUBAR flow meter was measured by a V ESSEN micro-manometer, and the flow Q_s computed by means of the formula (from the manufacturer's data sheet)

$$Q_s = K\sqrt{h_w}\text{litres/sec} \tag{6.1}$$

where $K = NSD^2\sqrt{\rho}$ depends on the flow meter diameter D and flow coefficient S, the air density ρ, and a conversion factor N. Values for these constants are given in Table 6.2.

Fig.6.7 shows the curve obtained by regression of the measured flow rate on to the A/D values from our flow meter. As shown, it is linear up to flow-rates of 4l/s. At higher flow rates, a quadratic correction term, obtained by a least squares match to the data, appears to match the deviation from linearity well. This quadratic relationship between A/D numbers and flow rate is implemented in the system software in order to provide actual values of flow rate.

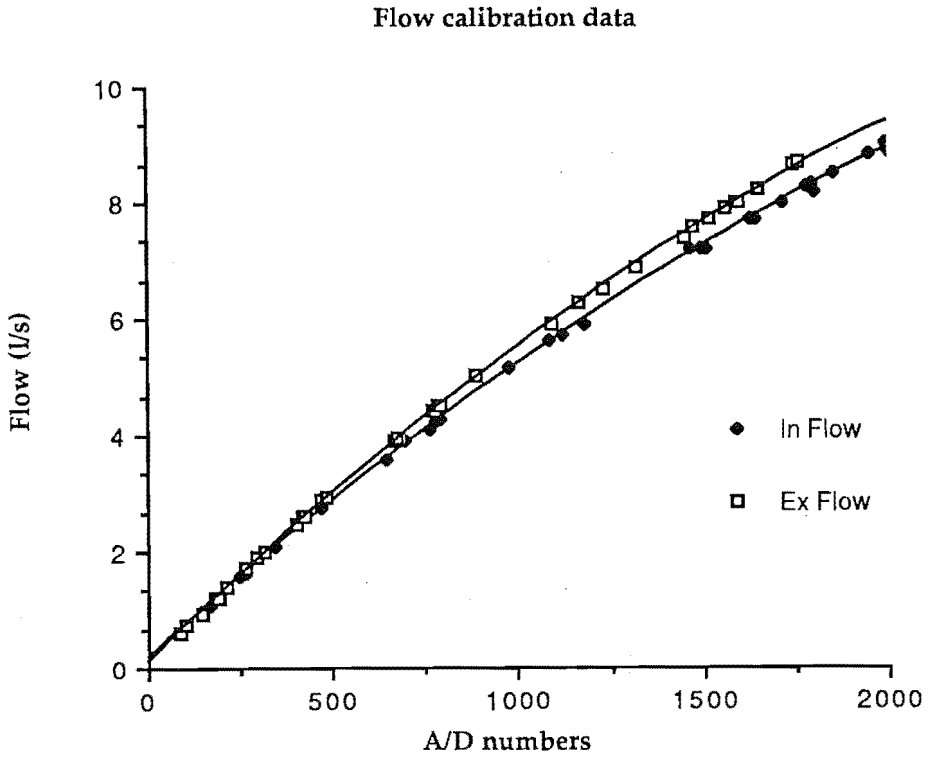


Figure 6.7. Calibration graph for the flow meter. Flow values were calculated via the formula (6.1) from measurements made with an ANNUBAR flow-meter. A-D values are given after subtraction of zero-flow offset. The quadratic regression curves are: Inhale flow

$$y = 0.18 + 5.728 \times 10^{-3}x - 6.72 \times 10^{-7}x^2$$

and exhale flow

$$y = 0.11 + 6.16 \times 10^{-3}x - 7.50 \times 10^{-7}x^2.$$

Note that the inhale flow and A/D numbers have been multiplied by -1 so that they can be graphed on the same axes as the exhale numbers..

6.3.3.2 The acoustic transfer function of the flow meter and facemask

The face mask and flow meter change the characteristics of sounds passing through them because they act as resonant cavities, thus attenuating some frequency components more than others. The flow meter can be considered to be a linear time invariant acoustic filter, with a transfer function $H_f(f)$ and impulse response $h_f(t) = \mathcal{F}^{-1}\{H_f(f)\}$. The filtered sound $s_r(t)$ is therefore related to the emitted sound $s_e(t)$ by

$$s_r(t) = s_e(t) \odot h_f(t). \tag{6.2}$$

It is straightforward to estimate $h_f(t)$ by Wiener filtering (Bates and McDonnell, 1986) once $s_r(t)$ and $s_e(t)$ are known.

The signals $s_e(t)$ and $s_r(t)$ refer to, respectively, a sound that is unaffected by the flow meter, and one that is so affected. In order to measure $s_e(t)$ and $s_r(t)$ through a single microphone, it was necessary to digitise them separately. I recorded a reference sound on a tape recorder, then replayed it twice, first digitising the sound directly and, second, digitising it after it had passed through the flow meter. The two digitised signals thus differed only in that one was affected by the transfer function of the flow meter. Sections of sound were extracted from each of the recordings by means

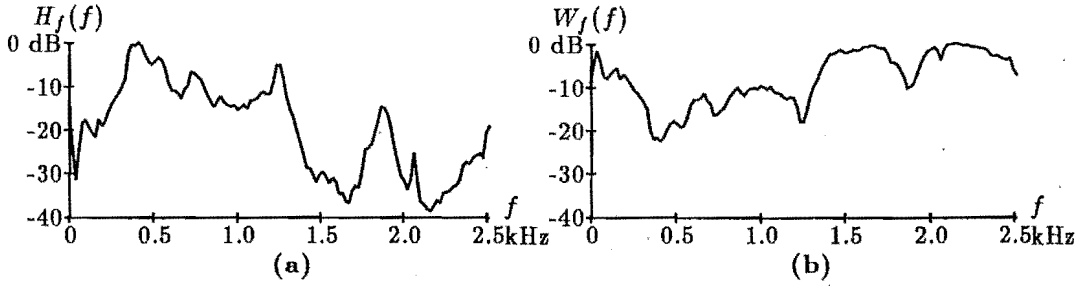


Figure 6.8. **a:** Acoustic transfer function $|\hat{H}_f(f)|$ of the flow meter (scale in dB). **b:** Estimated correction filter $|W_f(f)|$ for the flow meter transfer function.

of the "cough edit" program (§6.4). The position of the section from each recording was chosen so that the resulting signals, which I denote by $s_e(t)$ and $s_r(t)$ respectively, are time-aligned. Time alignment was easily achieved because the sounds exhibited a definite "starting" transient, which was positioned at $t = 0$ for both signals.

In order to reduce the effect of noise on the estimate $\hat{H}_f(f)$ of $H_f(f)$, I averaged several preliminary estimates together. Since the duration of $h_f(t)$ is short (of the order of a few ms) because of the dimensions of the flow meter, it is sufficient to divide $s_e(t)$ and $s_r(t)$ into segments of short duration and deconvolve each pair of segments separately. The resulting ensemble of filter estimates can then be averaged to produce $\hat{H}_f(f)$.

I divide $s_e(t)$ and $s_r(t)$ into segments of duration 51.2ms (256 samples), with the start-points of adjacent segments separated by 10ms (50 samples). Each segment is multiplied by a Blackman window and the FFT algorithm is employed to compute the Fourier coefficients. The magnitudes of these, for the m^{th} segment of $s_e(t)$ and $s_r(t)$ respectively, are denoted by $|S_e(f;m)|$ and $|S_r(f;m)|$. Note that this is the same procedure that I employ to compute the cough spectrograms (§6.4.4). In fact, I used the spectrographic analysis program described in §6.4 to compute the spectrograms of $s_e(t)$ and $s_r(t)$.

Each pair of spectra $|S_e(f;m)|$ and $|S_r(f;m)|$, for $m = 1..T/50\text{ms}$, where T is the duration of $s_e(t)$ and $s_r(t)$, is subjected to Wiener filtering to produce $|W_f(f;m)|$, the "spectrographic" estimate of the inverse filter of $\hat{H}_f(f)$. $|W_f(f;m)|$ is defined by

$$|W_f(f;m)|^2 = \frac{|\tilde{S}_e(f;m)|^2}{|\tilde{S}_r(f;m)|^2 + \phi} \tag{6.3}$$

where $\phi = 0.2$ is the Wiener constant and $|\tilde{S}_e(f;m)|^2$ and $|\tilde{S}_r(f;m)|^2$ are versions of $|S_e(f;m)|^2$ and $|S_r(f;m)|^2$, respectively, which have been scaled to have unit energy. The value of ϕ was chosen after examination of the results of several trial evaluations of (6.3) with a range of values for ϕ. The value of 0.2 for ϕ was the smallest that consistently avoided generating apparently spurious large values of $|W_f(f;m)|$ within the ranges of f and m wherein $|W_f(f;m)|$ was expected to be significant.

The final estimate of the inverse filter $|W_f(f)|^2$ is obtained by averaging each of the individual spectral lines in $|W_f(f;m)|^2$. Fig.6.8*b* shows a plot of $|W_f(f)|$, while Fig.6.8*a* shows the estimate of the flow meter transfer function $|\hat{H}_f(f)| = 1/|W_f(f)|$. It can be seen that $|\hat{H}_f(f)|$ contains pronounced spectral peaks and dips.

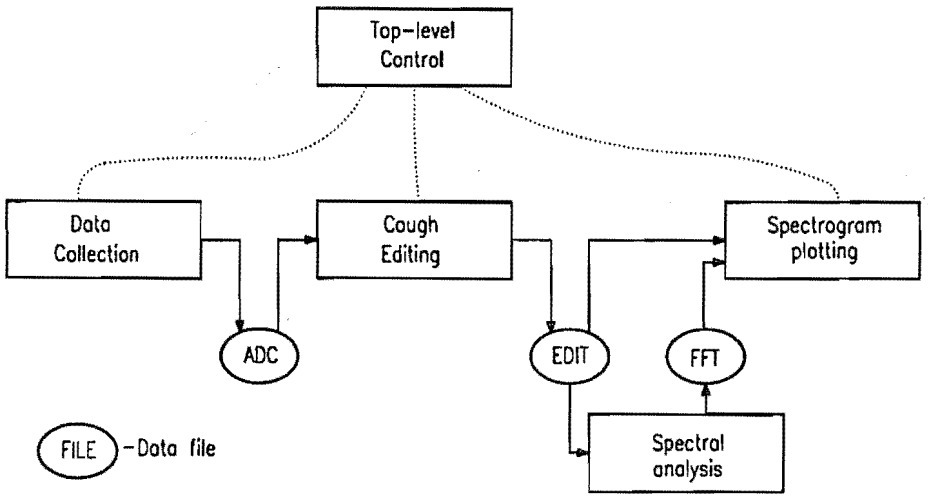


Figure 6.9. Block diagram of the cough system software, showing the various steps that are required in the analysis procedure. The flow of data between the various modules is also indicated. Each operational module accepts data from one set of data files and writes its output to another set of files.

6.4 System software

Fig.6.9 shows a block diagram of the COFF system software. The software consists of about 400kBytes of source code, written in the Modula-2 language (JPI TopSpeed). Graphics is implemented through the "MetaGraphics" interface (MetaWindow Corporation). In this section I describe the structure of the COFF software (§6.4.1), some details of its implementation (§6.4.2 through §6.4.3), and facets of its operation (§6.4.4).

6.4.1 Overview of COFF system software

The COFF program is written in a number of separate *modules*, each of which performs a particular part of the processing required in the system. Dividing the program into separate modules is necessary to ease the tasks of implementing and maintaining the software (cf. Brinch Hansen, 1977).

There are two type of modules in the program. *Operational modules* implement the blocks identified in Fig.6.9, being steps in the process of analysing the sounds. These are: digitising the sound and flow data, editing the data to isolate the coughs, performing spectrographic analysis on the cough sounds, and displaying and plotting the cough spectrograms. The *support modules* implement functions required in all parts of the program such as facilities for managing data files and interacting with the user.

The operational modules (refer to Fig.6.9) are functionally separate from each other. Data are transferred between each operational module via files on the hard disk. The program was structured in this way for three reasons. Firstly, the modules could be developed independently. This meant that clinical collection of cough sounds could begin as soon as the first module was completed. Secondly, it allows for development and expansion of the system in the future. New modules can be added without affecting existing modules. Thirdly, an individual module can be improved without affecting the overall system.

The different operational modules (refer to Fig.6.9) are each linked to, and controlled by, the *Top-level* module. Some of the modules require no human interaction,

and so operate automatically, while others are interactive. The interactive modules are activated by the selection of the appropriate menu entry in the top-level menu. The non-interactive modules, however, operate automatically after being activated at the start of program execution.

The non-interactive modules operate as "background tasks" in the multi-tasking environment provided by MODULA-2 (Wirth, 1983). Each background task is activated by the top-level module at the beginning of program execution. Thereafter it exists as a separate, concurrent, *process* from the main process. The processes can be treated as effectively executing in parallel (Brinch Hansen, 1977). Multi-tasking requires that a supervisory *scheduler* arbitrates between processes and ensures that each has a "fair" share of the CPU time (Brinch Hansen, 1977). The scheduler supplied with JPI Modula-2 is a *time-slicing* scheduler, in that it allocates to each process a "slice" of time to execute on the CPU. At the end of the time-slice, the scheduler *interrupts* the process that it is executing, and *swaps* it with another process. The scheduler maintains a list of all the processes, together with all the information necessary to restart the process when its turn arrives. Each process also has a *priority* associated with it, so that more important processes are allocated a greater proportion of the CPU time. In the COFF program, background processes are given a lower priority than *foreground* or interactive processes. This is to ensure that the program responds to the user without undue delay. The background processes only become active when the foreground process becomes inactive, such as when it is waiting for user input. At present, background processes are only employed to perform the spectrographic analysis and to plot spectrograms on the pen plotter.

Further details of the theory and design of multi-tasking programs can be found in many texts (cf. Brinch Hansen, 1977). The JPI TopSpeed Modula-2 manual (Jensen *et al.*, 1988) can be consulted for details of the particular scheduler utilised here.

6.4.2 Data management and control

In a project such as the cough sound research described here, large amounts of data must be efficiently managed (cf. Starmer *et al.*, 1987). The data that the COFF program must manage consist of the processed and unprocessed cough sound and airflow signals, replicated for each of the 30 or 40 subjects of the clinical study. Because of the modular way in which the COFF program is implemented(§6.4.1), data files exist for each of the intermediate stages in the analysis. In this section I describe how the files are managed so that the analysis can proceed without risk of confusing the data from one subject with that of another. Many texts on data management are available which may be consulted for background information on general techniques for organising extensive databases (cf. Inmon, 1981).

The flow of data and information between the various modules of the COFF program is indicated in Fig.6.9. Each stage in the analysis produces a set of output data files and associated index files. The index files contain information about the contents of the data files, as well as any information that may be supplied by the user at that stage.

6.4.2.1 Subject differentiation

The data files of each subject's coughs are kept separate by storing them in individual DOS sub-directories. Fig.6.10 illustrates this type of division. Sub-directories for the subjects Fred and Jane are contained in the directory DATA. A list of these sub-directories can be summoned by the user at the top level of the COFF program. Then

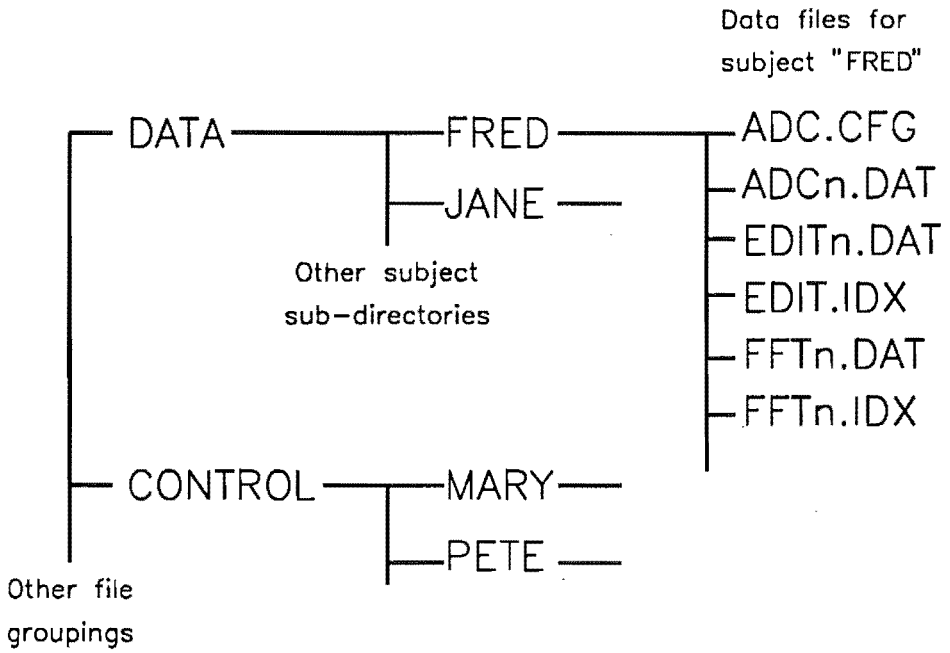


Figure 6.10. Data directory structure. Each subject has a sub-directory that contains all the data files pertaining to their coughs. The subject sub-directories may be grouped together into different "types" of subjects by putting each group into a different "data" top-level directory. The active sub-directory is chosen by the user via a menu at the top level of the COFF program.

one of them can be selected. Each of the interactive modules reads and writes data only to the currently selected sub-directory. This means that the user can select a subject, and then not have to worry further about file names or any of the other details of saving data to, or recalling it from, the disk.

As indicated in Fig.6.10, groups of subjects can be placed in different "parent" directories. This means that, for example, data from the control subjects can be grouped separately from that of the other subjects. The choice of which group is "active" is made by the user from the top level menu.

All the functions necessary to control the division between each subject's data files are implemented in a single support module. As mentioned above, the current subject is selected by the user in the top level of the COFF program. The complete path of that sub-directory is then stored in a file called `DATADIR.CFG`. This becomes the default until changed (even if the power is turned off in the mean time). Each of the interactive modules determines the current sub-directory by examining the `DATADIR.CFG` file. It then performs all its data file reading and writing to that sub-directory. This manner of communicating the name of the current patient directory to each module accords with the philosophy of having completely separate modules which communicate only via the hard disk (§6.4.1).

6.4.2.2 Data differentiation

Each subject's sub-directory contains data from the various stages of the analysis procedure, replicated for each of the six or seven coughs collected from that subject, and replicated again for the two sound channels and one flow channel. In order to keep track of all this data, several conventions are invoked. A "generic" file name is used to identify each type of data (i.e. raw data from the A/D routine, edited cough data,

File type	Contents	Format
ADCn.DAT	Raw data from A/D	Fixed size blocks of data + time flag for each block.
EDITn.DAT	Coughs selected by user in Edit module	Variable sized blocks. Each containing data from one cough. The length and location of each block is specified in the EDITO.IDX index file
FFTn.DAT	Spectrographic data of each cough	Variable sized blocks, each containing one spectrogram. Position and size of each block specified in the FFTn.IDX index file

Table 6.3. List of the files, and their generic names, used to store data at each stage of the COFF analysis.

Index file name	Contents
ADC.CFG	patient name, exercise time, sampling details
EDITO.IDX	Position and length of each block of data in the EDITn.DAT files. Also the location of the block in the ADCn.DAT files, and a comment entered by the user.
FFTn.IDX	Details on each spectrogram in the FFTn.DAT file. For each spectrogram: size of each spectral line, spacing between spectral lines, number of spectra, position and size (in bytes) of spectrogram in FFTn.DAT file.

Table 6.4. List of the index files used to record the positions and identities of the individual cough records within each data file.

spectrographic data). Each actual file name is composed of a generic name and a number identifying the signal channel. For example, the raw data file for channel number 14 is identified by the filename ADC14.DAT. Table 6.3 lists the generic file names used for each type of data.

The file for each type of data contains the relevant data from all the coughs of that subject. To identify each individual segment of data in a file, an *index file* is associated with each data file. Each of the index files contains information on the location and duration of each segment within the appropriate data file, together with identifying information such as the exercise time or any comments entered by the user. The index files and their contents are listed in Table 6.4.

In order to easily access data from any file, support modules are implemented for each type of data file. For instance, the module EditFiles implements procedures which allow a program to access any edited cough record, merely by specifying the subdirectory name, the channel number, and the cough record number. Similar facilities are also provided for the other types of data.

6.4.3 The human interface

The COFF program is used mainly by non-technical people during all aspects of data collection, management, analysis, and display. It therefore had to be simple and straightforward to operate. This is especially important for the parts of the program used in a clinical situation, where clinicians want to focus their attentions on the patient rather than the computer.

The "human interface" is a term which describes all those aspects of the program that affect how it interacts with the (human) user. This includes the methods of controlling the operation of the computer, the ways in which information is presented to the user, and the logical structure of the program. Brown (1988) presents a comprehensive coverage of human–computer interface design. He points out that the principal goal in designing the human interface is to minimize the effort required to perform any task. This naturally requires the program to be logically organised. The user can then easily "navigate" around the various functions of the program in the course of performing some particular operation (cf. Hogue and Fackrell, 1987). The effort required to perform any task is further minimised by avoiding "multi-level" menus whenever possible. This is in contrast to some other menu-driven programs where the user may need to select an item on a *top-level* menu, which causes a *sub-menu* to "pop up". The actual function is then selected from that menu or a further sub-menu (cf. Hogue and Fackrell, 1987). Such an approach may be necessary when a program has many functions available, but it means that accessing any particular function takes more effort than if they are all laid out in a single menu (Brown, 1988). A useful human interface on a computer program must not only be simple to operate, it must present information in a straightforward manner to the user (Brown, 1988).

The human interface of the COFF program was refined over a period of several months by a process of consultation with clinicians who were using a prototype program to collect coughs from patients.

The COFF program is logically divided into several smaller modules, each of which performs only the functions relevant to the particular task to which that module is dedicated. The different modules are organised in a "tree" type of structure, with each one accessible from the top-level of the program. The tree is "broad" rather than "deep", which means that nested sub-menus are avoided as far as possible. Deeply nested menus (where a particular function is accessed by sequentially selecting items on successive sub-menus) are confusing for the user because it is so easy to forget one's position in the "tree" (cf. Hogue and Fackrell, 1987). With a "broad" tree, however, one is never far from the top-level, so it is much quicker to move to another part of the program. One disadvantage with a broad menu tree is that a large number of menu options are required at the top level if the program contains many individual modules.

Within each module, every function is represented by an item on a single menu. The menu takes the form of a simple "control panel" on one side of the screen (Fig.6.11). Selection of an item on the menu is performed by means of a mouse or keyboard keypress. The use of the keyboard for typing in commands is avoided, although it is used for entering the patient's name. If so desired, the keyboard can be used for selecting items on the menu.

Information that the COFF program must present to the user consists of graphs of the sound and airflow waveforms, spectrograms of the cough sounds, information about the origins of the graphs being presented, messages about the status of the program itself, and the progress of the automatic data analysis. By employing a standard screen format in each of the different modules (see Fig.6.11), the user becomes accustomed to the way in which information is displayed, even in parts of the program which have not been used before.

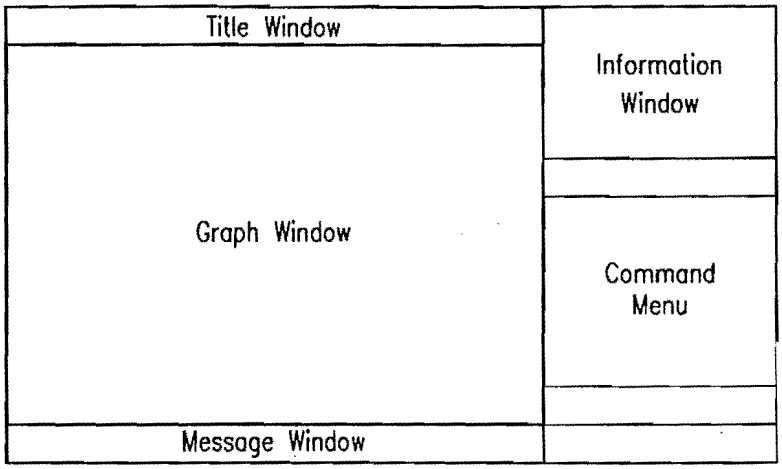


Figure 6.11. Standard screen format used in all parts of the system.

Each screen contains several *windows*. The main part of the screen consists of the *graphics window*. This contains the waveform or spectrogram that is currently being examined. On the right of the screen appears the *menu panel*. This has a separate item for each of the functions associated with the module. An item on the menu can be selected either by pointing to it with the *mouse* and pressing a *mouse button*; using the cursor-control keys on the keyboard to highlight the item and then pressing the <ENTER> key; or by pressing the key on the keyboard corresponding to the first letter of the word identifying the item. These different selection mechanisms are provided because different people have disparate preferences for controlling a computer. The menu panels for each module are laid out in a similar fashion with, for instance, "Help" and "Return to Main Menu" being the last two items on each menu.

The COFF program provides information to the user via two windows on the screen. The *information window*, on the top right of the screen, contains information about the data being examined. This includes things such as the patient's name, the time at which the current cough was collected, etc. The *message window* is an area at the bottom of the screen where messages from the program are displayed. These include warnings that something is amiss in the program (such as if the user tries to display data that the background process has not analysed yet), information messages about the operation of the analysis, and status messages from the program. The message window also prompts the user on the few occasions when something must be typed in (such as a comment to go with the edited cough data).

The fifth window on the screen is the title, which performs no purpose other than to identify which module is active. The display format of each of the modules is basically the same as described here, with appropriate modifications depending on the particular requirements of the module. The details of the user interface and its operation are discussed in §6.4.4.

6.4.4 Operational features of the COFF program

In this section I describe the operation of the COFF system, illustrating the discussion with examples of the screen display during a typical session with the COFF system.

In the "top-level" of the COFF program, the user can select a patient name from the list that is presented, or enter a new patient. After entering a new patient,

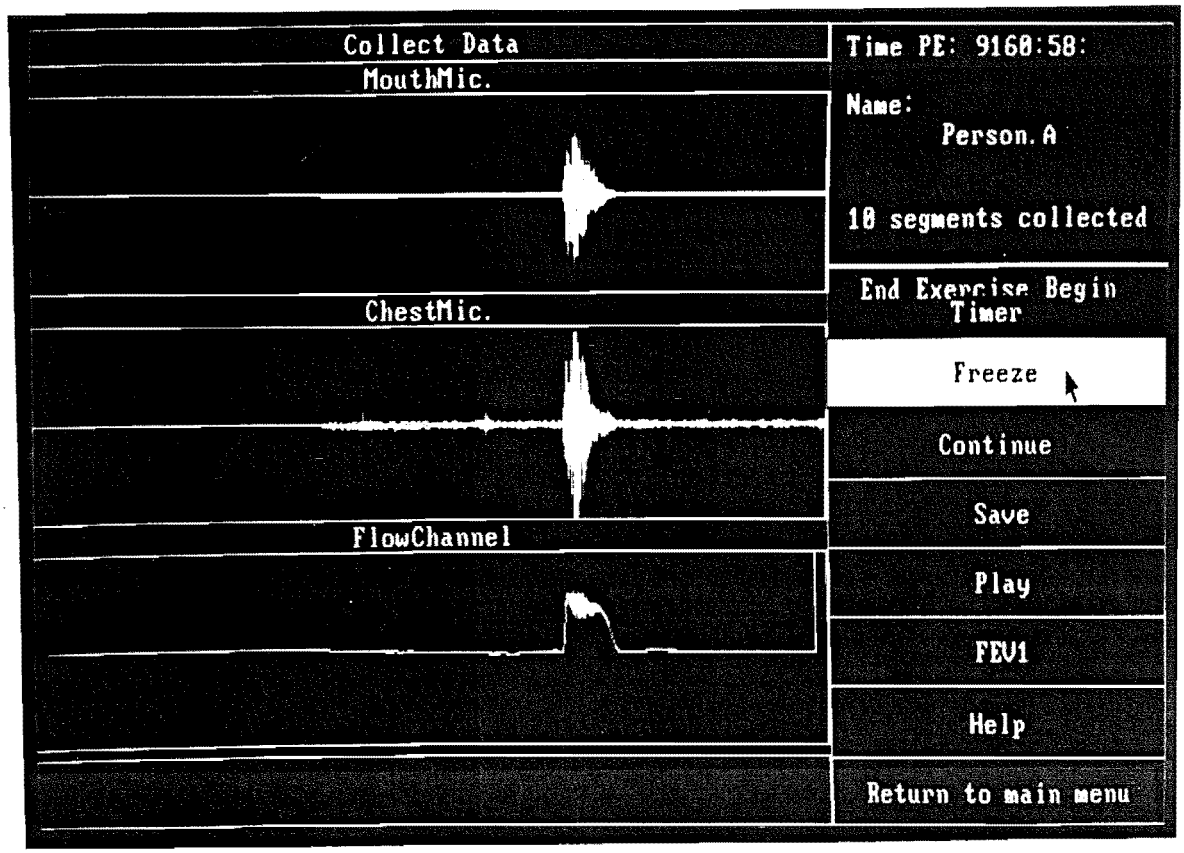


Figure 6.12. The screen of the COFF system when it is performing the "cough collection" task. See the text for details.

the user has the opportunity to change the sampling parameters to allow, for instance, single-channel collection from a tape recorder. Instead of a graph, the top level displays information about the currently selected patient, including the amount of data that has previously been collected and processed. Other menu entries at this level allow the user to select any of the "task" modules (described below), display some informative "Help" text, or to "Quit" from the program. If "Quit" is selected while the background processing (Spectrographic analysis or hard copy plotting) is in progress, the user is asked for confirmation before the program actually does terminate.

A photo of the screen as it appears in the "cough collection" module is shown in Fig.6.12. In the protocol for collecting cough sounds from a patient, coughs are collected at approximately two minute intervals after an exercise test. The data collection module samples the input signals continuously. When the patient actually coughs, the clinician presses a key to stop the sampling, retaining in memory the previous six seconds of sound. The sound can be replayed through a loud-speaker so the clinician can listen to the cough again. If satisfactory, the data can then be saved to disk, automatically including a label which uniquely identifies the time at which that cough was collected

Fig.6.13 shows a photo of the screen as it appears in the "cough edit" module. The three graphs represent, from top to bottom, the sound as recorded at the mouth and chest respectively, and the flow signal. The two vertical lines called "cursors" are used to identify the extent of a segment of the signals. The cursors are positioned by means of the "mouse". The segment so selected can be saved to the disk for (background) spectrographic processing, it can be replayed, or it can be re-plotted on an expanded time-scale ("zoomed") so that the details of the waveforms can be examined in greater

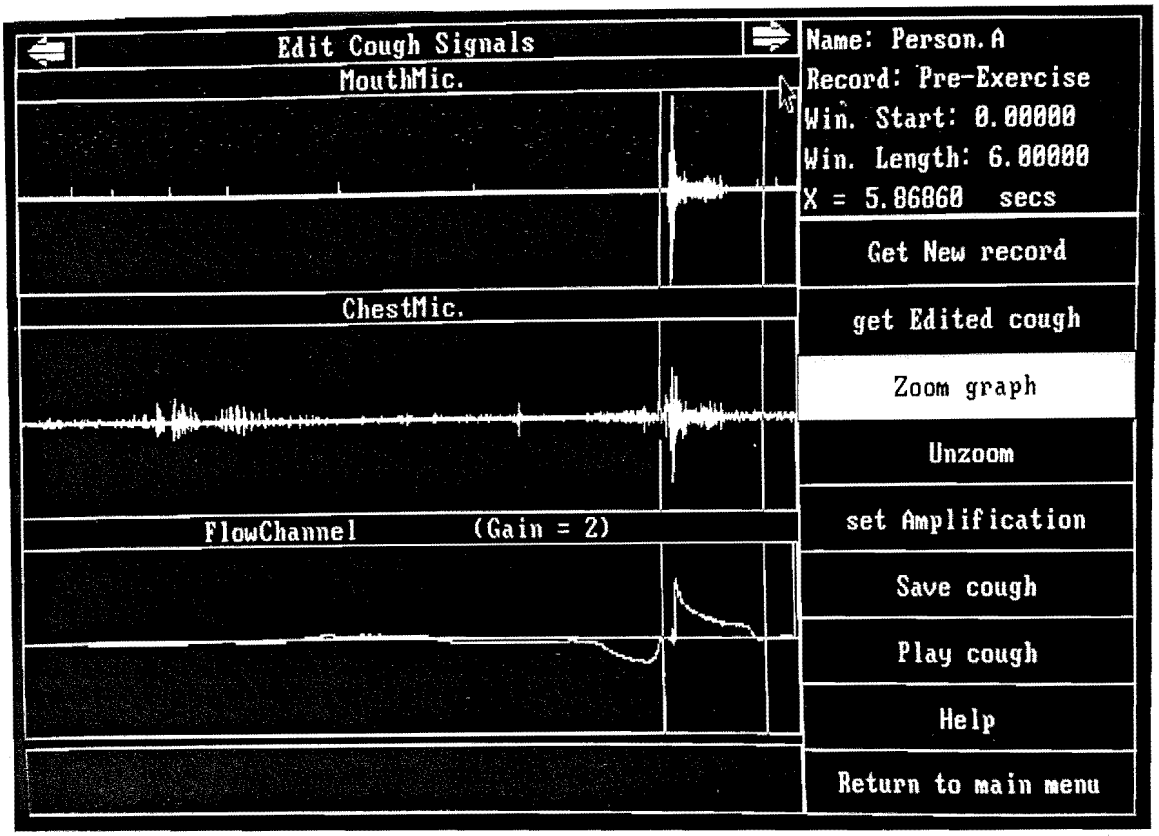


Figure 6.13. The screen of the COFF system when it is performing the "cough edit" group of functions. See the text for details.

detail. Each of these actions is represented by an entry on the menu, which appears on the right hand side of the screen (Fig.6.13). Note that any particular cough from either the raw data records or the edited cough records (see §6.4.2) can be selected and displayed, by means of the "Get new record" and "get Edit record" menu entries respectively.

The coughs that are selected and saved to disk by the "cough edit" module are automatically processed by the spectrographic analysis module, operating as a background process. Segments of 256 samples (51.2ms) duration are extracted at 50 sample (10ms) intervals throughout the cough sound. Each segment is then multiplied by a Blackman window (§1.3.1.1). Fourier coefficients are calculated by means of the FFT algorithm. The magnitudes of the first 128 Fourier coefficients (representing frequencies from 0 to 2.5kHz) are then computed. The spectrographic data are then saved in "FFT" files (see §6.4.2) for later display and any further analysis.

When the spectrographic data for a particular cough is available (typically each spectrogram is computed in about 1–5 minutes) it can be displayed. A typical screen display from the "spectrogram display" module is shown in Fig.6.14. The spectrogram is displayed as a "stack-plot" of spectral lines, with the sound and flow waveforms plotted alongside for reference. As in the cough edit module, the sound can be replayed through the speaker. In order to examine the spectral features in more detail, any particular spectral line can be selected by "pointing" to it with the mouse and clicking a mouse button. The spectral line then appears in the small window near the top of the screen (Fig.6.14). In addition, the frequency, epoch, and magnitude of the selected point are displayed in the small information box to the lower right of the

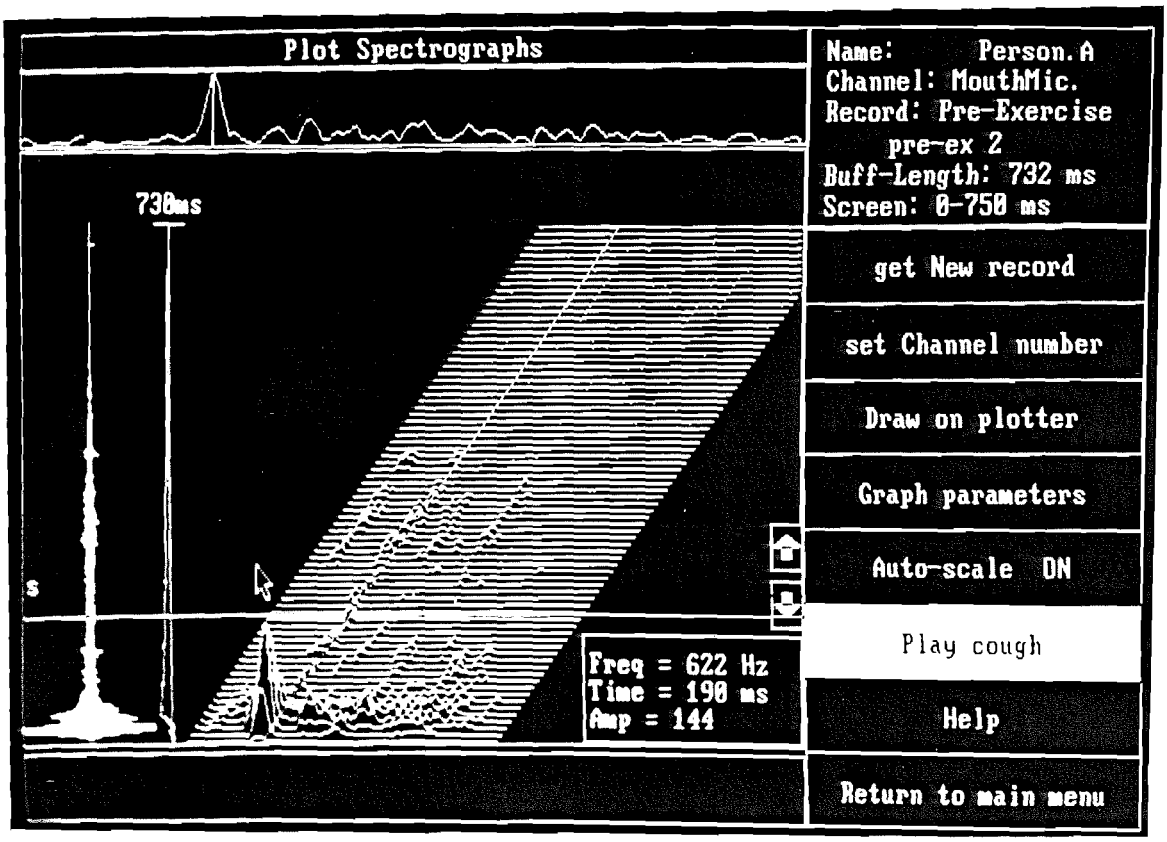


Figure 6.14. The screen of the COFF system as it appears when displaying a spectrogram of a cough. See the text for details.

spectrogram itself. The spectrogram of any particular cough from the current patient can be displayed by selecting it via the "get New cough" menu entry. The format of the spectrogram plot can be altered by means of a sub-menu accessed via the "Graph parameters" entry on the menu shown in Fig.6.14.

Spectrograms can be plotted on paper to produce a permanent record of the cough. The plotter is controlled by a sub-menu that is accessed from the "spectrogram display" module. Typical spectrograms produced by the COFF program are shown in Fig.6.15.

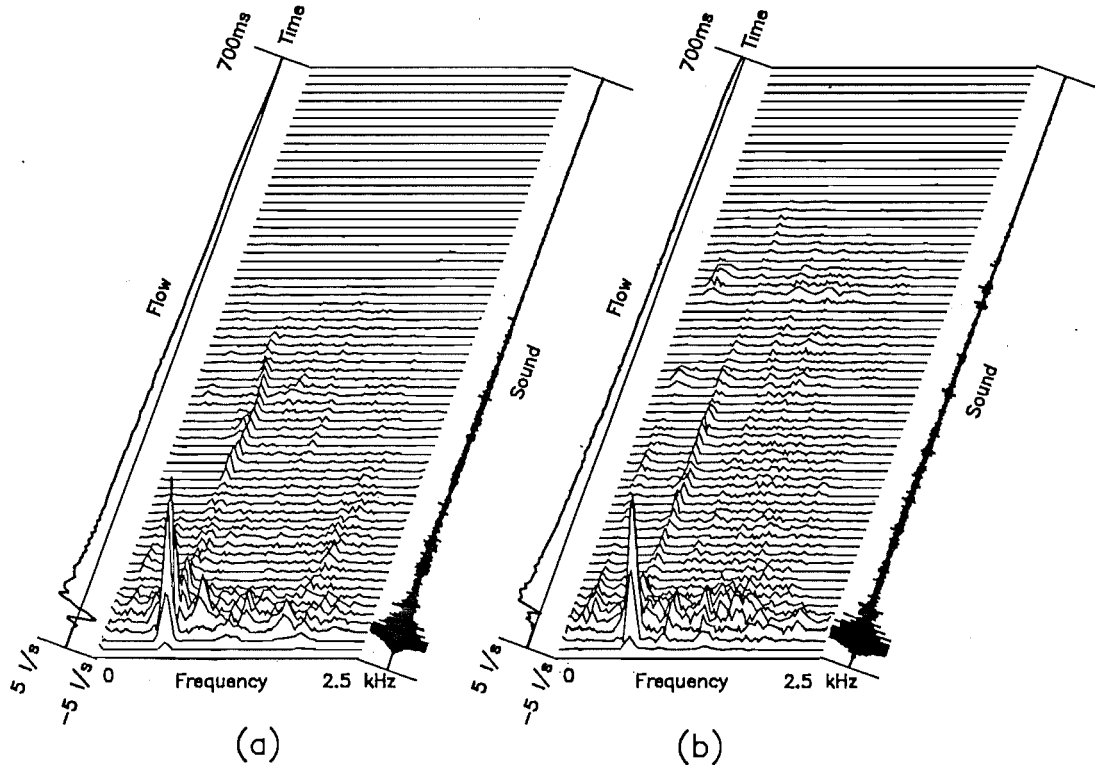


Figure 6.15. Spectrograms produced by the COFF program from a non-asthmatic child. The spectrogram on the left is of a cough sound recorded at rest, while the one on the right is of a cough recorded six minutes after the completion of an exercise test. Note that some text has been added to the figure by a graphics editor program.

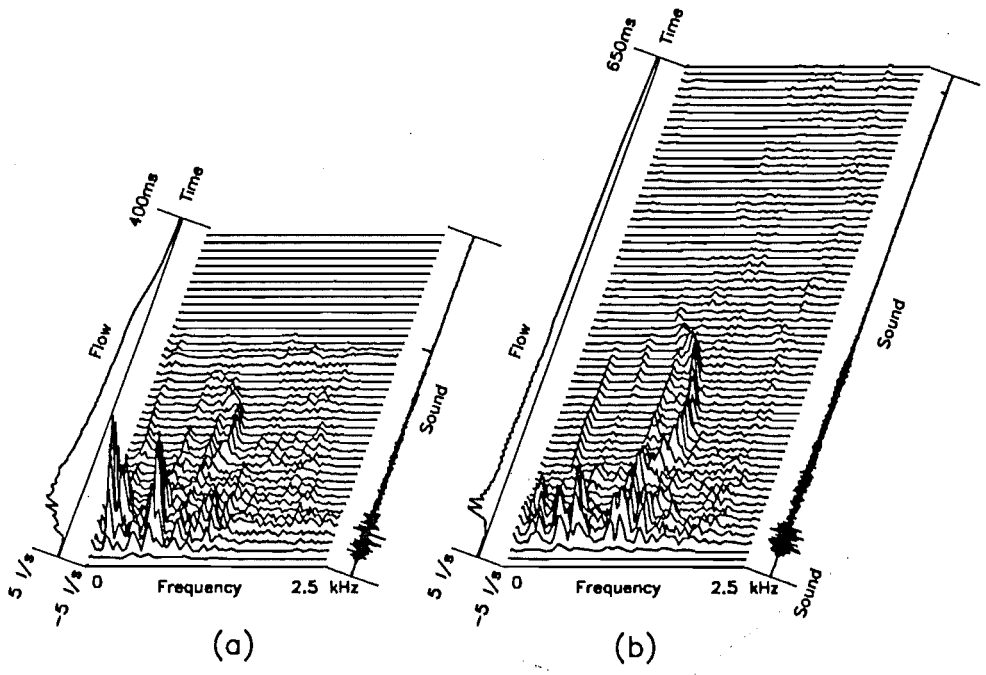


Figure 6.16. Spectrograms produced by the COFF program from coughs by an asthmatic child.

Chapter 7

Analysis of Hector's Dolphin vocalisations

Hector's dolphin, *Cephalorhynchus hectori*, is a small coastal dolphin native to New Zealand (Baker, 1978). As part of a comprehensive study of these dolphins (Slooten and Dawson, 1988), many sounds were recorded from free-ranging individuals. The purpose of the studies described in this chapter was to derive a procedure that quantitatively characterise the recorded sounds using signal processing techniques.

The research reported here was performed in close collaboration with Steve Dawson, as part of his Doctoral studies in the Zoology Department. Many of the results could not have been obtained without this inter-disciplinary approach. Although there was close consultation between us during all parts of the analysis, the data collection, described in §7.2.1, was undertaken entirely by Steve Dawson. The analysis in §7.2.2 through §7.2.2.5 and §7.3 was performed largely by myself, and the statistical analysis of §7.2.3 mainly by Steve. I report this research from my viewpoint as an engineer, so that the conclusions that I make are largely related to the engineering aspects of the study. Alternative perspectives, stressing the biological and behavioural implications of this work are given by Dawson and Thorpe (1990) and Dawson (1990) respectively.

In §7.1 I present a brief introduction to sonar systems in general, the use of echo-location and other vocalisations by marine mammals, and the techniques that have been employed in past studies to examine animal sounds. I also provide some background on Hector's dolphins, and give an overview of the physiology of delphinid sound production. In §7.2 I describe the techniques by which descriptive features were extracted (automatically) from the recorded sounds. The statistical techniques used to describe the variation in the vocal repertoire of the dolphins are also described. Finally, §7.3 describes the analysis techniques used to evaluate the echo-location performance of the sounds.

7.1 Background

7.1.1 Hector's dolphin

Hector's dolphin, *Cephalorhynchus hectori* is native to New Zealand, but is closely related to three other species, namely *Cephalorhynchus commersonii* (which chiefly inhabits the eastern coasts of South America), *Cephalorhynchus eutropia* (which is found in Chilean waters), and *Cephalorhynchus heavisidii* (inhabiting Southern African coasts). Hector's dolphin is the world's smallest dolphin, and one of the rarest (Dawson and Slooten, 1988). Its behaviour and ecology has only recently been examined in detail, by Slooten and Dawson (1988). The analyses of Hector's dolphin vocalisations

reported here are a continuation of these studies. Much of the work described in this chapter is also reported in publications by Dawson and Thorpe (1990), Thorpe and Dawson (19XX) and Thorpe *et al.* (19XXb). The behavioural implications of the sounds made by Hector's dolphins, as analysed in the ways described in this chapter, are discussed by Dawson (1990).

Hector's dolphins live in small groups, and seldom stray more than 8km from shore (Dawson and Slooten, 1988). They feed on many types of small fish and squid, from the surface down to the sea floor (Slooten and Dawson, 1988). Like many other dolphin species, they are inquisitive and investigate any (slow-moving) boat that enters their area (Slooten and Dawson, 1988).

In common with many other marine mammals, the incidental catch of Hector's dolphin in gill-nets (nets which are left in place for a period of time, and which catch fish as they swim through them by entangling in their gills) is of considerable concern (Slooten and Dawson, 1988; Dawson, 19XXa). This problem is especially acute around the Banks Peninsular area because of the substantial gill-netting effort in the region. Dawson (1988) estimates that upwards of 228 dolphins were killed in both commercial and recreational gill-nets between 1984 and 1988 in this region. A seasonal ban on gill-netting around Banks Peninsular was enforced from 1989 in order to protect the remaining population of some 660 individuals (Dawson, 19XXa). In light of the demonstrated echo-location abilities of odontocetes, it is natural to ask whether Hector's dolphins are capable of detecting monofilament nylon nets (Dawson, 19XXb). Part of the motivation for carrying out the research described in this chapter was to investigate whether the emitted sounds are, in principle, such as to permit the dolphins to perceive nets.

The sounds emitted by Hector's dolphin (or any of the other species in the *Cephalorhynchus* genus) have not been studied in detail previously. The only previous reports of recordings of the sounds of Hector's dolphins are by Watkins *et al.* (1977) and Dawson (1988). The sounds of the closely related species *Cephalorhynchus commersonii* have been examined by (among others) Kamminga and Wiersma (1982), and those of *Cephalorhynchus heavisidii* by Watkins *et al.* (1977). Hence it is interesting to analyse Hector's dolphin's sounds and compare them with those reported of other odontocetes. This was another part of the reason for undertaking the research reported in §7.2 and §7.3.

7.1.2 Sonar systems

Echo-location is the technique of actively probing the environment with some sort of radiating signal and determining the location(s) and form(s) of object(s) in the environment by analysis of the signal echoes. Most commonly, the radiation employed to probe the environment is electromagnetic (used in radar systems by aircraft, ships etc, cf. Skolnik, 1980) or acoustic (used in sonar systems by ships and some animal species cf. Kinsler *et al.*, 1982).

The range R of a *target* can be calculated from the estimated time delay t_d between the transmitted and received signals:

$$R = \frac{t_d}{2c} \tag{7.1}$$

where c is the wave velocity of the radiation in the medium. The factor 2 in (7.1) occurs because the signal must travel to the target and back again. The angular position of the target is determined by transmitting only a narrow beam of radiation, and using it to scan through each radial segment of the surrounding environment. The velocity at which the target is travelling relative to the sonar system can be calculated by measuring

the *Doppler* frequency shift in the echo. Doppler occurs because the reflected waves are compressed (or expanded) by the motion of the target towards (or away from) the sonar system. Hence the frequency of the echo is scaled by an amount relative to the (radial) velocity of the target (Skolnik, 1980, Chapter 3). The Doppler shift Δf for a single frequency component f is (Kinsler *et al.*, 1982, p416)

$$\Delta f = 2\frac{\dot{R}}{c} f \tag{7.2}$$

where $\dot{R}$ is the radial component of the velocity difference between the source and the target.

The resolution to which an echo-location system can determine the range and velocity of a target is determined by the form of the signal employed. Unfortunately, the requirements of good resolution in both range and velocity of the target are in opposition to each other. A short wide band signal (ideally, an impulse) is required for accurate range resolution, but a long duration (narrow band) signal (ideally, a signal of infinite extent) is required for accurate determination of target velocity. Hence the type of signal employed for any application must be suited to the type of information required. The *ambiguity* of a particular sonar signal describes its joint range and velocity resolving capabilities (cf. Woodward, 1953; see §7.3 of this thesis).

The resolution to which angular position can be determined is restricted by the directionality of the transmitting and receiving transducers. Since directionality is approximately proportional to transducer size (relative to the signal wavelength), human engineered sonar systems usually employ signals of the smallest wavelength that is feasible after taking into account signal propagation factors and other pertinent practicalities (Skolnik, 1980, Chapter 2).

The maximum range at which a sonar system can detect targets is limited by the signal-to-noise ratio of the received echoes. This depends on the transmitter power, signal attenuation in the medium, reflectance of the targets and the amount of noise in the environment (Kinsler *et al.*, 1982, §15.8). For underwater sonar, signal attenuation varies with frequency. Higher frequency sounds are attenuated to a far greater extent than those of low frequency. However, high frequency sounds are advantageous for several reasons: directional transducers are physically small and hence more realistic; interference from other sound sources is limited to those in the near vicinity; wider bandwidth signals are easier to produce and detect; and environmental noise levels are lower than at lower frequencies (Kinsler *et al.*, 1982, pp412–414).

Echo-location was first observed in bats and marine mammals about the same era that technological sonar and radar systems were being developed (cf. Griffin, 1979; McBride, 1956; Watkins and Wartzok, 1985). In addition to bats and some species of odontocetes (whales, dolphins and porpoises), echo-location capabilities have been observed in several other animal species such as swifts and shrews (cf. Henson and Schnitzler, 1979).

7.1.3 Vocalisation and echo-location by marine mammals

Marine mammals collectively make a wide variety of sounds, ranging from short "clicks" to musical "whistles", "grunts", and "burst-pulses". Each species, however, has its own repertoire of vocalisations, which may include one or more distinct types of sounds (such as those mentioned above). Sounds are employed by animals for many purposes, such as communication between individuals or groups, warnings, etc (Seyfarth *et al.*,1980a,1980b). In addition, a few species of odontocetes (toothed whales) have demonstrated the ability to echo-locate by means of sonar "clicks" (Evans, 1973). Although it is sometimes assumed from this observation that click-type sounds are solely

employed for echo-location , while other types of sounds have alternative uses (cf. Popper, 1980), this is not an unequivocal distinction. For example, Watkins (1979) reports observations of sperm whales emitting click-like sounds in a manner that strongly suggested that they were not using them for echo-location . In addition, some odontocete species only emit click-type sounds (Watkins, 1979) and it is inconceivable that such vocal and social animals do not employ their vocalisations for some form of communication (Herman and Tavolga, 1980).

The characteristics of the clicks produced by marine mammals differ, in some cases widely, between species (Evans, 1973). Most clicks are less than 1ms in duration, with some species emitting clicks shorter than 50μs (Evans, 1973). The spectral characteristics of the clicks also vary widely. For instance, *Tursiops truncatus* emits clicks of broad bandwidth, with significant energy in a range of 100Hz–100kHz (Evans, 1973). By contrast, the clicks of the *Cephalorhynchus* genus are of much narrower bandwidth, about 20kHz, with the energy centred at a frequency of about 120kHz (cf. Kamminga and Wiersma, 1982; this chapter, §7.2.3.1). The sounds of the larger whales are of correspondingly lower frequency than those made by smaller cetaceans such as dolphins. For example, the largest cetaceans, blue and fin whales, emit sounds whose energy extends from about 200Hz down to less than 20Hz (cf. Payne and Webb, 1971; Watkins and Wartzok, 1985).

In addition to the variability between species, there is a great variability in the characteristics of clicks reported in different studies of animals of the same species. For example, the clicks of *Inia geoffrensis* presented by Evans (1973) are very wide-band, while Wiersma (1982) reports clicks from an animal of the same species that are relatively narrow-band. Whether these differences are due to different experimental procedures or due to differences between the individual animals is not clear (see Diercks *et al.* (1973), Watkins (1974), Au (1979,1986), and §7.1.4.1 of this chapter for further comments on possible sources of variability in delphinid sounds).

Whistles consist of long (up to several seconds) sequences of musical sounding tones (Popper, 1980). Some studies have noted the occurrence of dual component sounds, with both click and whistle sounds present simultaneously (cf. Dziedzic, 1978, pp49–69). However, whistles have not been observed from Hector's dolphins or any of its near relatives (Dawson, 1988; Evans *et al.*, 1988).

Further details of the sounds produced by marine mammals, and the uses to which they put those sounds, can be found in texts such as those by Popper (1980) or Herman and Tavolga (1980). Evans (1973) and (Watkins and Wartzok, 1985) review the different types of cetacean click-type sounds that have been recorded.

7.1.4 Physiology and models of sound production and perception

Fig.7.1 shows a cross-section of a dolphin head and indicates the various structures associated with sound production and perception. §7.1.4.1 gives an overview of what is currently understood about the mechanisms by which dolphins produce sounds, while §7.1.4.2 does the same for how they perceive sounds.

7.1.4.1 Sound production

The exact mechanism by which sounds are produced in cetaceans is the subject of some debate (Norris, 1969, p404). It is possible that there may be two separate sound production mechanisms, for whistles and clicks respectively (Dziedzic, 1978, pp49–52).

Because the sound producing organs are deep within the animal's head, it is difficult to investigate the mechanisms by which sounds are generated. Cetaceans have no vocal cords. It appears that (at least for echo-location) sounds are produced by the

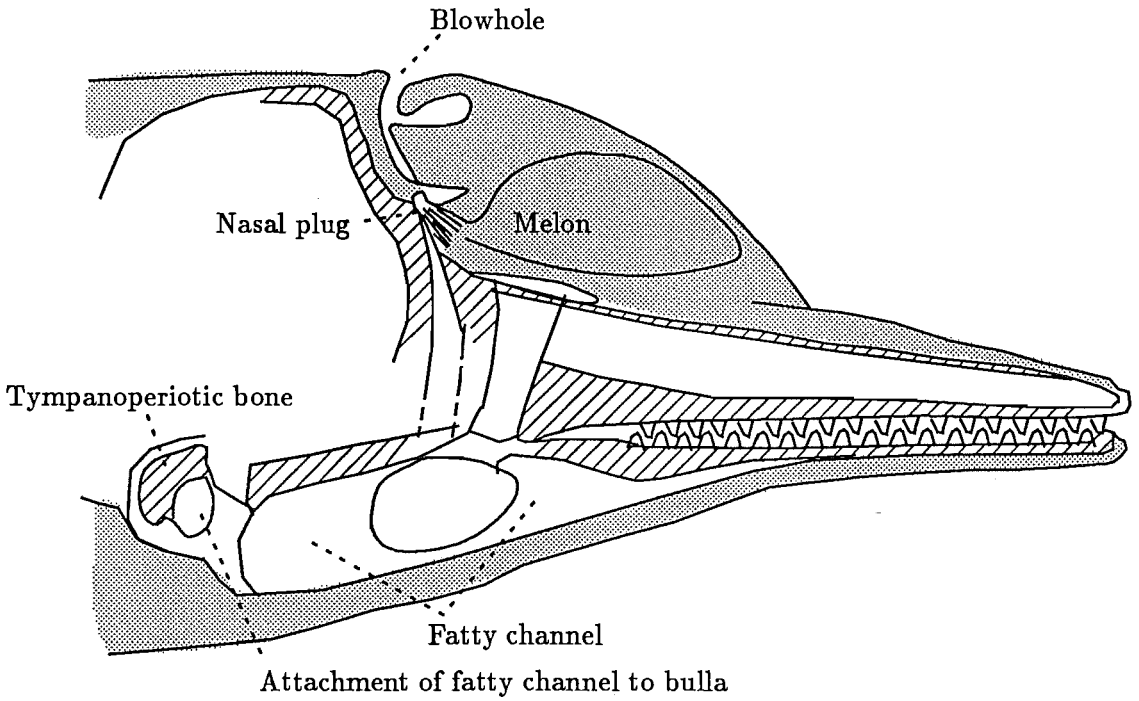


Figure 7.1. Section of a dolphin head, showing the various structures associated with sound production and perception (after Popper, 1980).

"nasal plugs" (Popper, 1980; Cranford, 1988). These are hard muscular organs which jut into the anterior surface of the nasal passage as it exits from the skull (Fig.7.1). Ridgway and Carder (1988) have shown that the intra-nasal pressure increases when the animal is phonating. Furthermore, when they introduced a leak into the nasal passage (by means of an open catheter), sound production was seriously impaired. Several studies have shown that the spectral character of the high frequency click sounds does not change when the animal is breathing a helium air mixture (cf. Cranford, 1988). This indicates that these sounds are not produced in an air-filled resonant cavity (as are human speech sounds), but rather by the vibration of some tissue structure (Cranford, 1988).

The spectral content of some low frequency whistles change when the animal breathes a helium air mixture (Cranford, 1988), which supports the hypothesis that there are two sound producing mechanisms. It seems reasonable to assume that the click-type sounds are produced by the nasal plugs and associated organs, while other sounds (such as whistles) are produced in another structure such as the larynx (cf. Norris and Evans, 1966).

The click sounds emitted by dolphins are very directional (Au *et al.*, 1986). Several researchers have postulated that the *melon*, a fatty, oval-shaped tissue situated in the bulbous forehead of the animal (see Fig.7.1), acts as an acoustic lens (Popper, 1980; Cranford, 1988). The melon abuts onto the nasal plugs (Cranford, 1988) and it seems very likely that sounds generated by vibration of the nasal plugs or some associated organ are coupled directly into the melon.

The waveforms and spectra of recorded clicks vary widely according to the orientation of the dolphin's "sound beam" with the sound transducer Au *et al.* (1986). This variability could partly explain some of the variation in click waveforms observed

by different researchers (§7.1.5). Au (among others) postulates that the sound beam is formed, not only by refraction within the melon, but also from *internal echoes* off structures such as bones and air sacs in the animal's head. Kamminga and Wiersma (1981) suggest that it is these internal reflections that often cause click pulses to exhibit phase discontinuities (see Fig.7.6).

Dziedzic (1978) and Popper (1980) both review cetacean sound production mechanisms in more detail than is presented here.

7.1.4.2 Models of sound perception

Like the mechanisms of sound production, the perception of sounds by dolphins is still something of a mystery. Physically, the ear canals are quite small, so that it is probable that sound is coupled from outside the animal to the ears through other structures, such as the lower jaw (Norris, 1969). The lower jaw contains a fat-filled canal which terminates near the middle ear (Popper, 1980, p29). This canal has a lower acoustic index than water so could act as a slow-wave antenna. The "jaw-hearing" hypothesis was supported by an an experiment which indicated that a dolphin could hear better when its jaw was not covered by an acoustic hood (Brill *et al.*, 1988).

There are several models of how a dolphin analyses echoes in order to obtain target information. One model incorporates a *matched filter* detector, which is often employed in technological echo-location or communication systems. This correlates the incoming signal with a copy of the transmitted pulse. A large correlation value indicates that an echo of that pulse is present in the signal, while a small correlation indicates the probable absence of a pulse echo (Cook and Bernfield, 1967). Such a detector provides the optimum capability for detecting the presence of a specific signal in background noise (Woodward, 1953). However, evidence from several recent studies indicates that matched filter detection is probably not employed by echo-locating mammals (cf. Altes, 1988; Au and Moore, 1988).

Au and Moore (1988) have performed an experiment to investigate the type of click detection employed by dolphins. They instrumented a computer to process clicks emitted by a dolphin so that simulated echoes could be returned to the dolphin. By varying the echoes that they returned, they could observe the effect of different click characteristics on the ability of the dolphin to detect the echoes. Specifically, they investigated the effect of returning closely spaced multiple copies of a click to the dolphin. If two or more clicks were contained within an interval of less than 260μs, they found that the detection performance of the dolphin improved in proportion to the number of clicks within that interval. However, there was no improvement if the clicks were spaced by more than 260μs. These results support the hypothesis that the dolphin is using a simple *energy detector*, with a time constant of 260μs, to detect the pulses. An energy detector estimates the presence or otherwise of a desired signal in background noise on the basis of the energy envelope (§3.1.1) of the signal plus noise. The signal is deemed to be present when the envelope exceeds a certain threshold. The time constant refers to the interval over which the energy of the signal is integrated. Hence if more than one click occurs within an interval equal to the time constant, the energy envelope is proportionally increased and the probability of detecting the presence of the echo is improved. By contrast, the detection performance of a matched filter detector does not depend on the spacing between individual clicks (Cook and Bernfield, 1967, Chapter 2).

Altes (1988) comes to a similar conclusion after examining the echo-locating performance of bats. He proposes a *spectrogram correlation* model of detection, in which the matching is performed on the time-varying spectral magnitude of the sounds. As described in §3.3.1, the spectrogram can be understood as a set of energy envelopes

corresponding to the outputs of a bank of filters. Therefore, this model of perception is basically the same as that proposed by Au and Moore, with the refinement that the signal is filtered before its envelope is computed. This model is similar to the filter-bank model of the cochlear that is invoked to explain human auditory perception (cf. §2.2.2).

A consequence of the energy detection model is that the envelope of the signal is particularly significant. When a rapid sequence of clicks is heard, a tone equal to the repetition rate is perceived. This so-called *time-separation pitch* (TSP) arises because the spectrum of the click train has a "rippled" nature, with peaks occurring at harmonics of the repetition rate. TSP occurs even if the click "train" only consists of two clicks. Due to non-linearities in the detection process, the "fundamental" can be perceived as a tone (see §2.2.2.2). Hammer and Au (1980) propose that dolphins determine the structure of a target (at least partly) by means of the TSP induced by the multiple echoes from the target. Recent experiments by Au and Pawloski (1989) have indicated that dolphins can detect TSP between 2 and 75kHz, which corresponds to inter-click delays of 500–13μs. Experiments where echoes from (dolphin) clicks are slowed-down and played to human listeners have indicated that they are able to discriminate the echoes from different targets as well as the dolphins can (Au, 1988a). The listeners' comments indicated that they discriminated between the targets on the basis of the time-structure of the echoes.

7.1.5 Studies of cetacean vocalisations

Some studies of cetaceans have concentrated on the characteristics of their sounds which are pertinent to echo-location (cf. Evans, 1973; Kamminga and Wiersma, 1981; Au, 1988a), while others have investigated the animals' actual echo-location abilities (cf. Dziedzic, 1978; Murchison, 1979; Au, 1988b). Studies which focus on the characteristics of the sounds have as one of their goals the elucidation of the mechanisms by which the dolphins are able to extract information about their environment from the sounds. The models which are proposed to explain the production and perception of sounds by dolphins are introduced in §7.1.4.

Most studies of the *echo-location abilities* of cetaceans have been performed on Atlantic bottlenose dolphins *Tursiops truncatus* or Harbour porpoises *Phocoena phocoena* (Kamminga, 1988). These creatures have demonstrated the ability to detect wires or monofilament nylon of about 0.2mm diameter (Norris, 1969; Evans *et al.*, 1988), to discriminate between spheres that are only 10% different in size (Popper, 1980, p35), and sheets of metal that vary in thickness or composition (Norris, 1969, p417). The range at which Bottlenose dolphins are able to detect targets has been variously estimated at about 75m for a steel ball 7.6cm in diameter (Murchison, 1979) to about 1km for detecting a school of fish (Evans, 1973). Although the low frequency sounds of fin whales have been observed at great distances (trans-oceanic), it is unclear whether these sounds are employed for echo-location (cf. Payne and Webb, 1971).

Another group of studies on cetacean sounds is related to the behavioural correlates of various vocalisations. Such studies have usually focussed on the audible types of cetacean sounds (cf. Lilly and Miller, 1961; Ford and Fisher, 1982). The reason for this is partly technological, because standard audio recording equipment is readily available and much less expensive than wide-band equipment. Also, many researchers seem to have assumed that the high frequency clicks are used strictly for echo-location . However, as mentioned in §7.1.3, there seems to be no reason why such signals could not also be used for communication.

Researchers conducting behavioural studies often assign recorded sounds to various categories (Clark, 1982; Chabot, 1988) and then attempt to determine whether different categories can be associated with particular behaviour patterns (cf. Herman

and Tavolga, 1980; Clark, 1982; Sjare and Smith, 1986). Elucidation of the communication systems of cetaceans is of considerable interest in view of the animals' large brains (cf. Morgane *et al.*, 1986) and complex behaviour (cf. Connor and Norris, 1982). However, acoustic repertoires have been analysed and described in detail for only a few cetacean species.

Just as the recording of animal sounds has been limited by available equipment, so too has their analysis. The tendency in the past has been to classify animal sounds according to the subjective impressions of human listeners (cf. Dreher, 1966). The development of electronic instruments which display visual representations of the time and frequency components of sounds (e.g. the spectrogram, see §3.3.1) has allowed researchers to compare sounds pictorially. Sound spectrograms have been compared subjectively (cf. Marler and Peters, 1981) and quantitatively. There have been several approaches to the problem of quantitatively comparing visual images of sounds. Many researchers have manually extracted parameters directly from the images (cf. Dawson and Jenkins, 1983; Clark, 1982). Others have digitised printed images of the sounds (cf. Miller, 1979; Chabot, 1988). In either approach, the resulting data are usually analysed using multi-variate statistical methods (cf. Sparling and Williams, 1978). In only a few recent studies have sounds been digitised directly on to a computer before being analysed with the aid of digital signal processing techniques (cf. Goedeking, 1983; Clark *et al.*, 1987).

7.2 Characterisation of acoustic repertoire

This part of the research had two aims. The first was to codify whatever correlations there might be between sounds made by and the observed behaviour of Hector's dolphins. An essential preliminary to this was to quantitatively characterise recorded sounds in order to classify them into different types. The second aim, which was reached through the process of achieving the first, was to investigate the application of digital signal processing techniques to the analysis of animal vocalisations. As stated in §7.1.5, researchers in this field have only recently begun employing sophisticated signal processing techniques. With the advent of affordable and powerful computers, coupled with sophisticated signal processing software, it is becoming increasingly feasible to automate the extraction of features from recorded sounds, thus relieving much of the drudgery of research in this field (cf. Davis, 1986).

Fig.7.2 summarises the approach I have adopted to recording, analysing, and characterising Hector's dolphin sounds. The sound recording procedure is described in detail in §7.2.1, while the methods used to analyse the sounds and extract characteristic features from them are presented in §7.2.2. The statistical analysis of the resulting feature data-set is described in §7.2.3. Finally, §7.2.4 assesses the results and discusses some of the difficulties encountered during the extraction and statistical analysis of the features.

7.2.1 Sound recording procedures

Recordings were made in Akaroa Harbour (43°50'S; 172°56'E), and in the nearby inshore waters of the south coast of Banks Peninsula, New Zealand, over the summer seasons of 1986/87 and 1987/88.

Sounds were recorded with a BRÜEL AND KJÆR 8103 hydrophone, BRÜEL AND KJÆR 2635 charge amplifier, and a RACAL STORE 4DS recorder operated at a tape speed of 60ips (152.4cm/s) or 30ips (76.2cm/s). At these tape speeds, the recorder has a signal/noise ratio of 40dB and a minimum frequency response of 300Hz

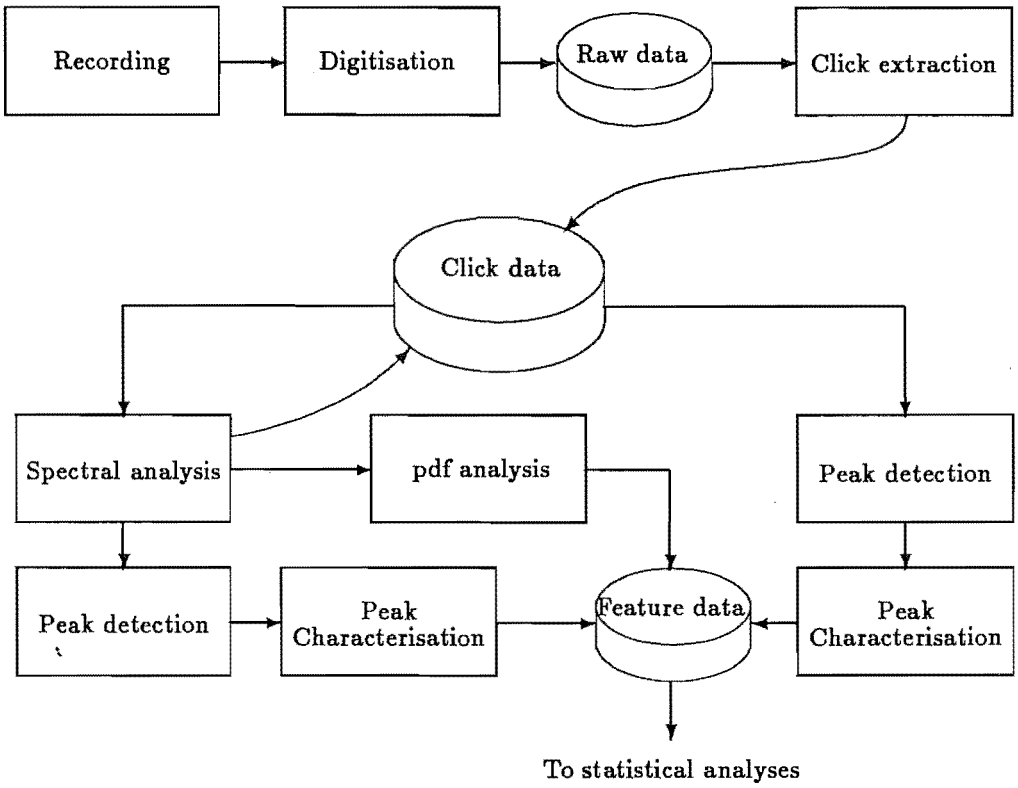


Figure 7.2. Block diagram illustrating the extraction of descriptive features from the sounds.

to 150kHz(+/−3dB). A BRÜEL AND KJÆR 4223 hydrophone calibrator was used to generate a reference level from which received sound pressure levels could be calculated. Behavioural notes were dictated onto another channel of the recorder, together with timing signals and the settings of the recorder's signal input amplifier, which were varied to avoid saturation. Recordings were made in calm conditions (wind speed <10 knots) from a four-metre inflatable boat.

Twelve hours of sounds were recorded, and subsequently all the tapes were transcribed at reduced speed in order to establish an index of their contents. The location of sounds that corresponded with known behavioural contexts were identified and listed. All tapes were transcribed at 1/16 speed or less using a MULTIGON INDUSTRIES UNISCAN II spectrum analyser or GOULD OS4000 digital storage oscilloscope to view the signals. Signal output was also fed to a NAGRA IV-L tape-recorder operating in 'test' mode so that its calibrated modulometer could be used to read relative sound pressure levels. The transcripts provided a directory of the sounds, their location on tape, relative sound pressure level, and a written version of the commentary.

From the transcripts, sounds that were made in known behavioural or biological contexts were located. Click sequences that contained clicks from more than one dolphin were discarded. At a tape replay speed 1/32 of the original recording speed, eight second segments of each sequence were digitised at a sampling rate of 20kHz with a DEC LPA-11 12 bit A/D converter. To avoid aliasing (§1.2.5.5), the signals were filtered with a 48dB/octave low-pass filter (KEMO VBF/8), which had a 3dB cutoff frequency of 9kHz, before they were digitised. The effective sampling rate was thus 640kHz and each digitised segment corresponded to 0.25 seconds at the original recording speed. Each segment, or *record*, was stored on magnetic tape for later analysis.

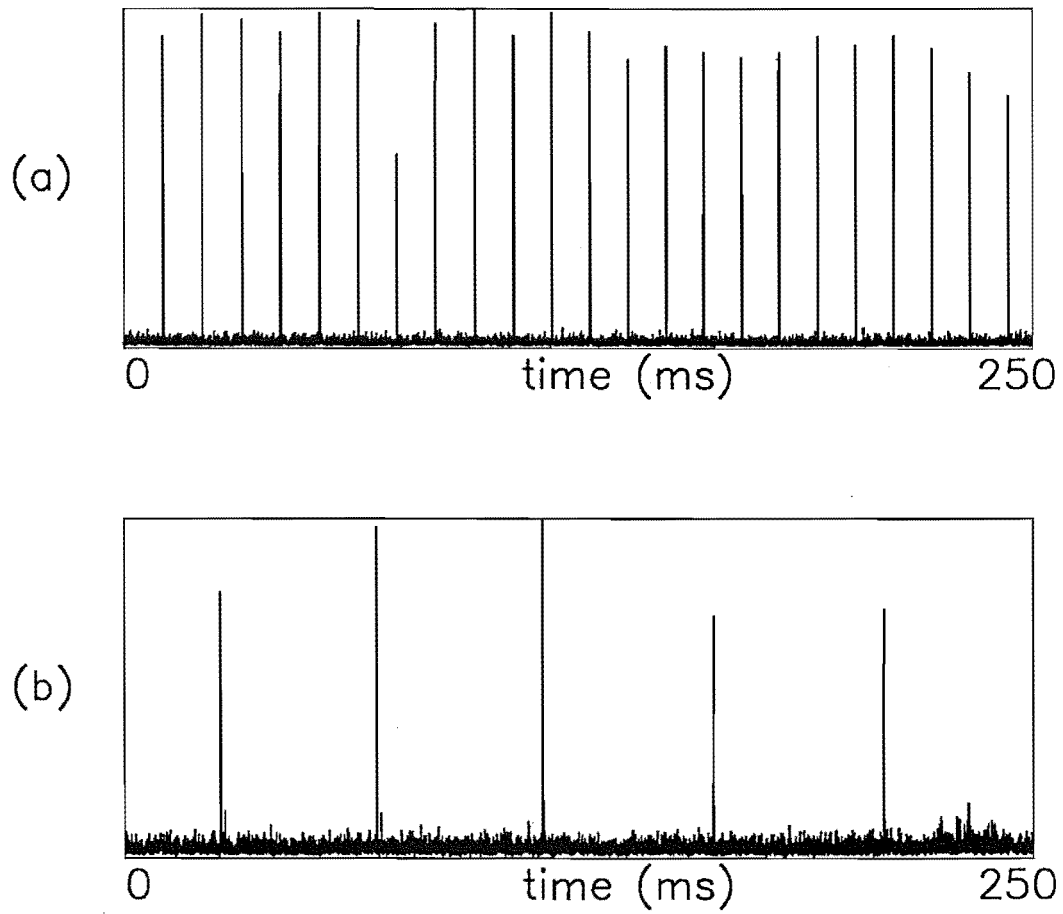


Figure 7.3. Envelope profiles for two representative sound records. **a:** 10 ms and **b:** 50 ms inter-click interval.

7.2.2 Characteristic features

In this section I introduce and describe the features that were employed to characterise the sounds. §7.2.2.1 presents representative examples of the sounds that were encountered during the study, while §7.2.2.2 introduces a set of features that describes the time and frequency domain structures of the "click" type of sounds (see §7.2.2.1). §7.2.2.3 describes the procedure for extracting features from each 0.25 second long sound record. §7.2.2.4 explains how synthetic click envelopes were reconstructed from the estimated feature variables, to assess how well the features characterised the clicks. Finally, §7.2.2.5 comments on the automatic implementation of the feature extraction scheme.

7.2.2.1 Examples of sounds encountered

Examples of the sounds that were encountered during the study are shown in Figs.7.3 to 7.5. In Fig.7.3 the envelope profiles for two of the sound records are depicted. The (RMS) envelope was obtained by the method described in §3.1.1 (see §7.2.2.3 for details). The interval between clicks varied from about 1ms to 160ms (median = 27ms) in the available sound recordings.

As indicated in Fig.7.3, the dolphin sounds consist of very short "clicks", fairly regularly spaced within each record. In order to examine the features of the clicks more

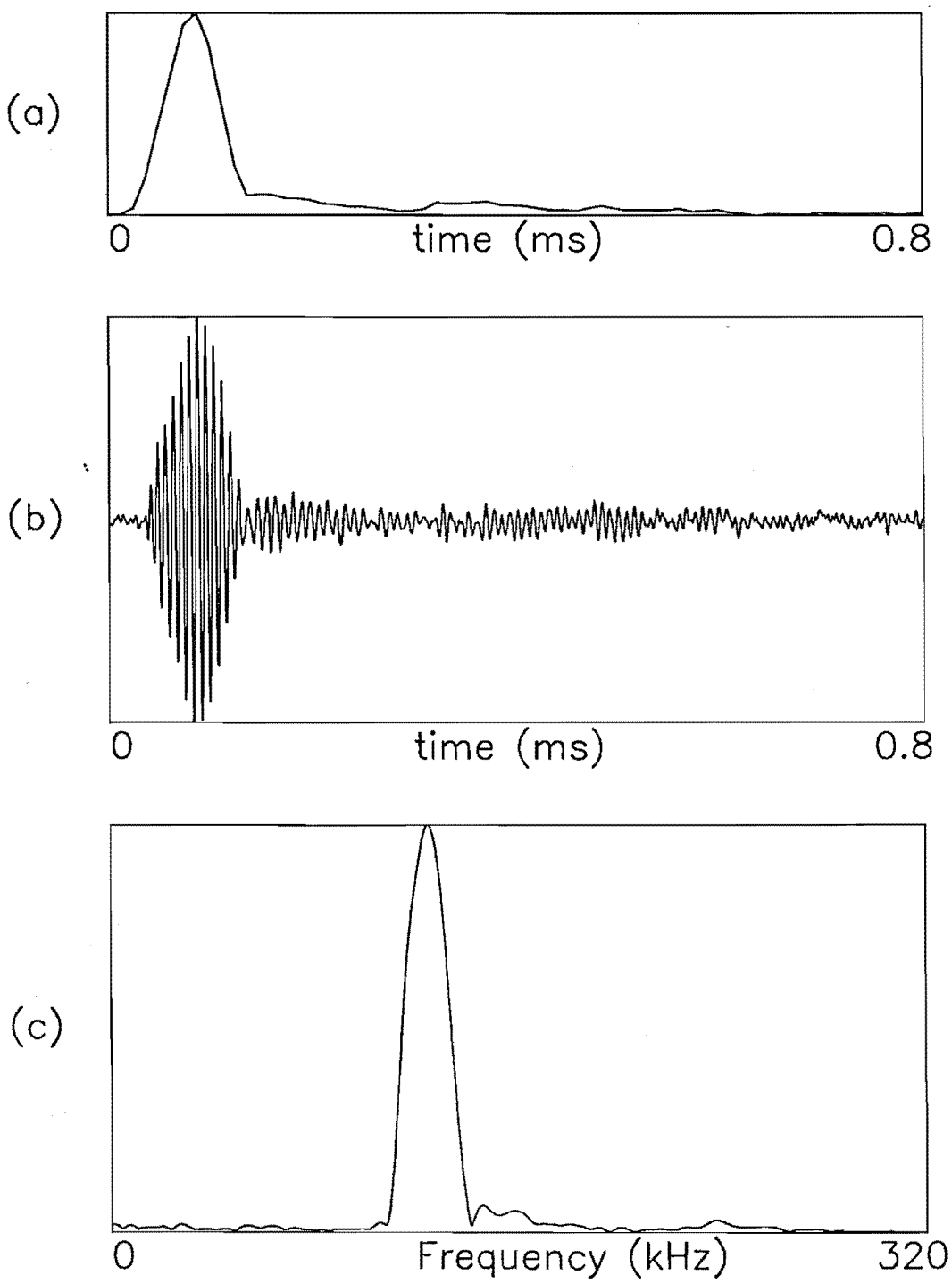


Figure 7.4. **a:** Average envelope profile, **b:** click waveform and **c:** average spectrum for a narrowband click with a single click in the time domain.

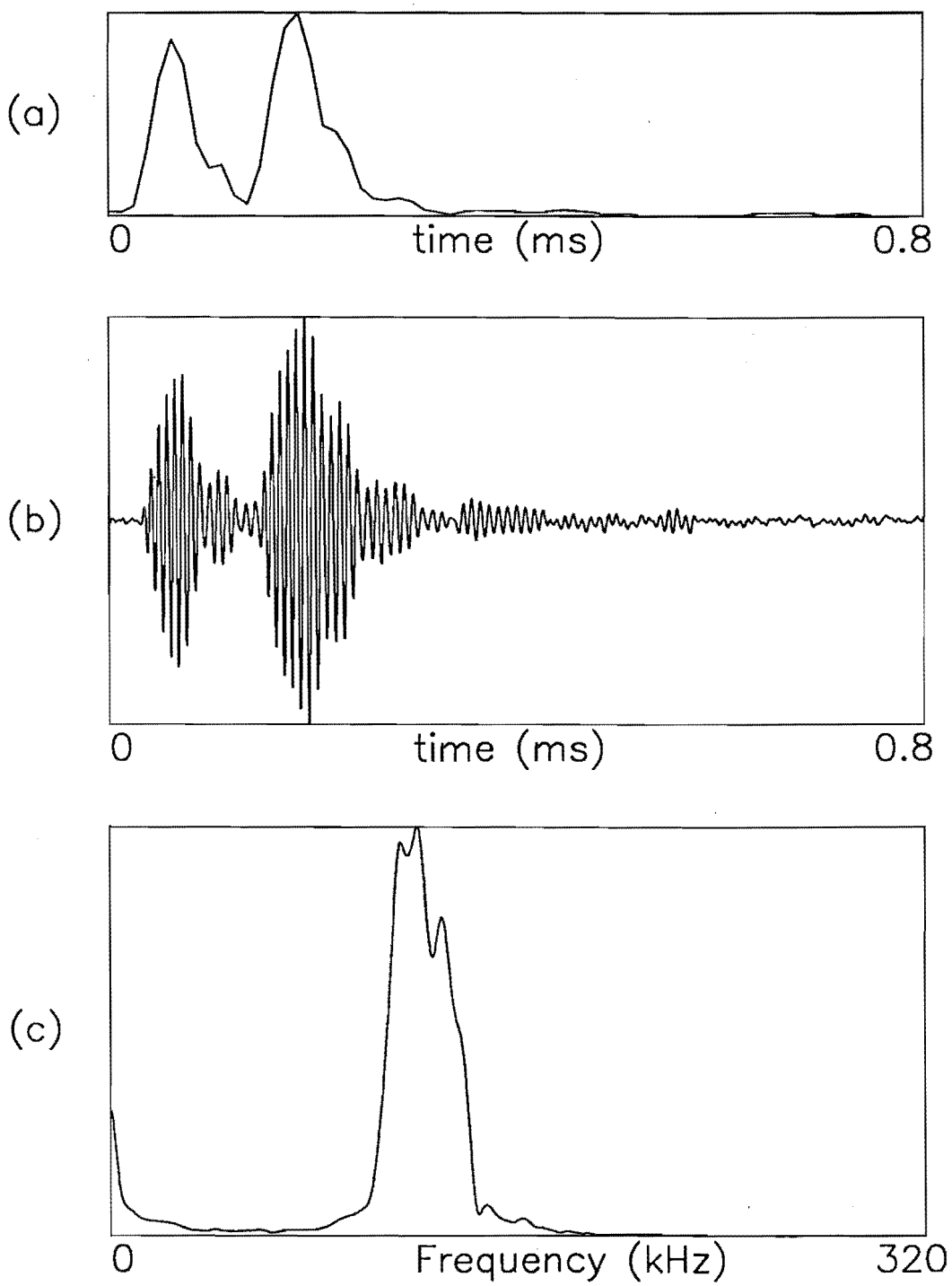


Figure 7.5. **a:** Average envelope profile, **b:** click waveform and **c:** average spectrum for a click with a double click in the time domain and "double" spectral peak.

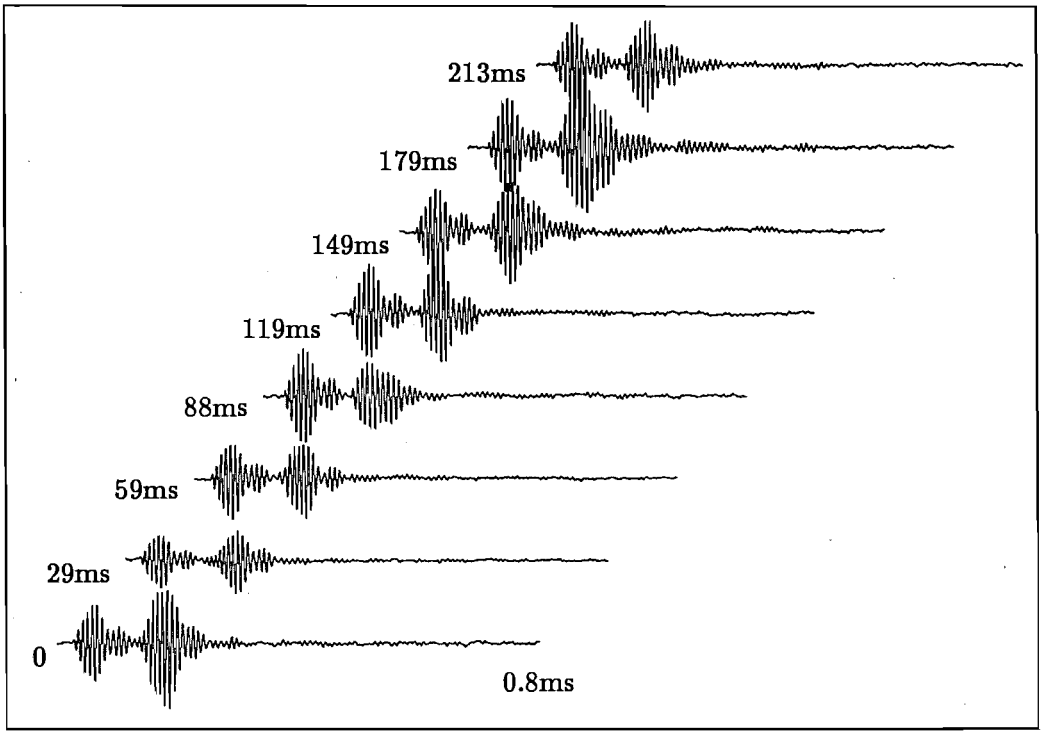


Figure 7.6. A series of clicks from a single record. Each trace is 0.8ms in length, and the start of each one relative to the whole record is indicated by the number on its left.

minutely, individual clicks were extracted from each record. The waveforms of typical clicks are shown in Fig.7.4 and 7.5, together with their envelope profiles and plots of their spectra. Note that the clicks are short, narrow-band signals with a dominant frequency of approximately 125kHz. Each click envelope consists of one or more peaks spread over 100–600μs.

7.2.2.2 Descriptive features of click waveforms and spectra

This section qualitatively describes the features which I chose to characterise the dolphin clicks. §7.2.2.3 details the steps in the procedure for extracting these features from the sound records. The complete list of variables which are invoked to characterise the clicks within any particular sound record are itemised in Table 7.1. Note that I use the term "feature" to indicate a characteristic of the clicks that I wish to describe, and the term "feature variable" to mean a particular quantity that can be given a value characterising a certain aspect of a feature. Three general features of the clicks are evident from Figs.7.3 through 7.5.

The first feature consists of the inter-click interval, which varies widely between records, and may increase or decrease within a record, as is also commented on by Au (1979). Hence I estimate the average inter-click interval for a record, together with the minimum, maximum, variance, and average trend of the inter-click intervals throughout a record (Step 3 in the procedure described in §7.2.2.3). The latter four variables characterise the amount of variation of the inter-click interval during the record.

The second feature is the envelope profile of each click. This can be affected by echoes and multi-path distortion and must be cautiously interpreted. However, by averaging the envelope profile for each click in a record, the effects of echoes are minimised, since they vary from click to click and hence tend to cancel out. Note that

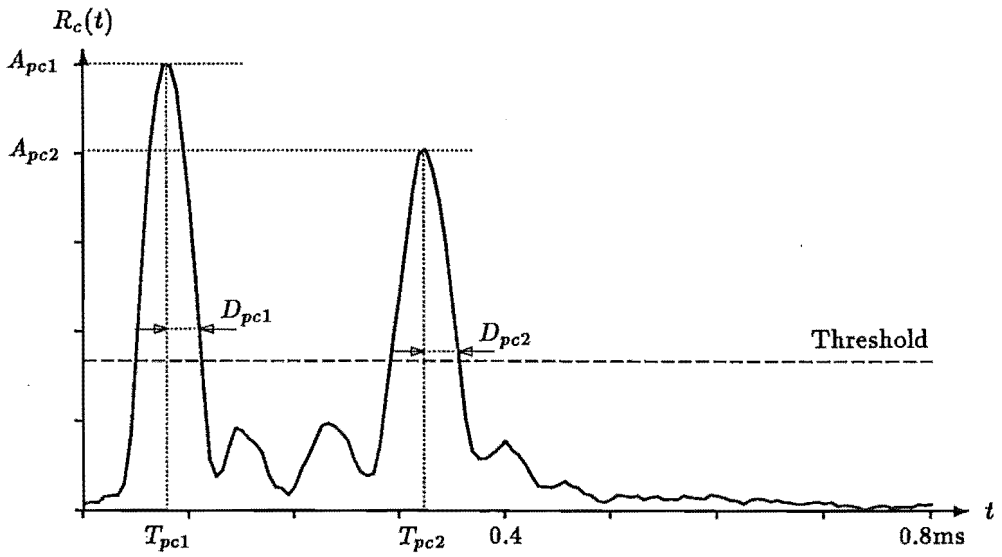


Figure 7.7. Diagram showing the features abstracted from the average click energy envelope. The amplitude, position, and decay-time of each peak greater than 1/3rd the maximum were determined.

even though the the length of each record is only 1/4 second, there can be considerable variation in the individual clicks, as is illustrated in Fig.7.6. The shape of the envelope profile is described by quantifying the amplitude, width (or decay time), and position of each of the individual peaks (Fig.7.7). Only those peaks that were greater than 1/3rd of the maximum amplitude were considered. If there were more than 4 such peaks, only the largest 4 are employed as feature variables. Step 4 of the procedure (§7.2.2.3) defines these variables. The duration of each click is also estimated by considering the click envelope as a probability distribution function (pdf, §1.3.3) and computing its standard deviation (Step 5)

The third feature comprises the major characteristics of the spectrum of each click. As stated above, most of the signals were narrow-band. A simple way to characterise such signals is by their *dominant* frequencies (defined in Step 8 of the procedure described in §7.2.2.3) and half-power bandwidths. Signals with complicated spectral shapes (see Fig.7.5) can be characterised by the amplitudes, centre frequencies and half-power bandwidths of several of the larger peaks in their spectra (Fig.7.8). As with the envelope features, only the 4 largest peaks, having amplitudes greater than 1/3rd of the maximum peak, are considered. Step 9 of the procedure (§7.2.2.3) describes how these variables were estimated. The average spectrum can also be characterised by treating it as a pdf and estimating the mean and standard deviation of the spectral energy distribution (Step 10).

In order to determine if the frequency content of the clicks varies from click to click, I estimate the peak frequency for each click. I then compute the average, standard deviation, and average trend of the peak frequency throughout each record (Step 7). For the purpose of comparing different methods of characterising the dominant frequency component in a click (i.e. that single frequency which best characterises the click spectrum), I also estimate the average zero crossing rate in each click, as described in Step 8, computing the same statistics as mentioned above for the peak frequency.

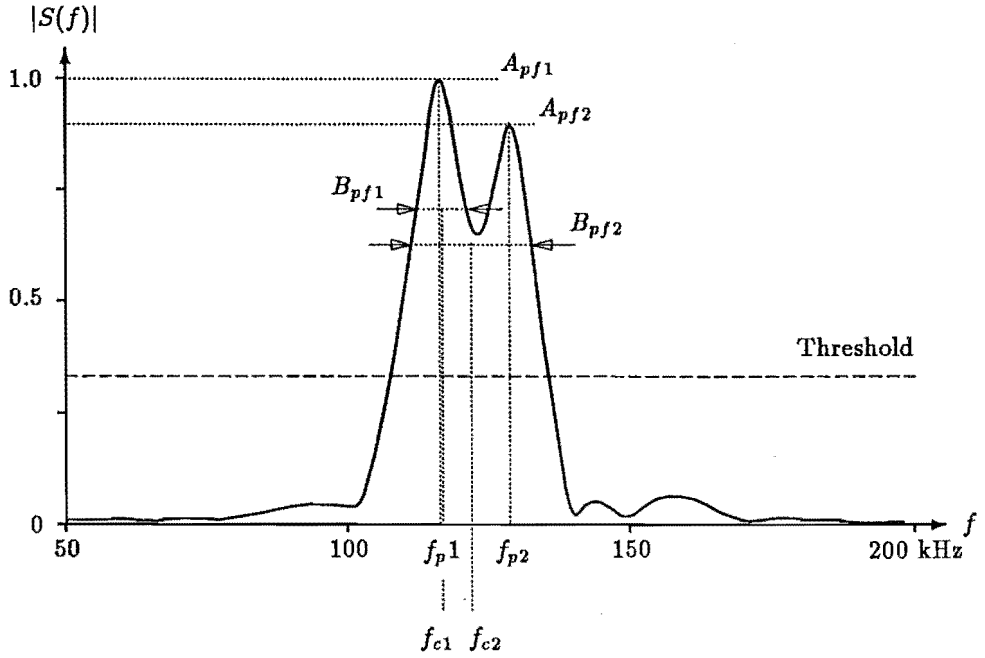


Figure 7.8. Abstraction of features from the average spectral magnitude. For each peak (up to 4) greater in amplitude than 1/3rd of the maximum peak, the amplitude, frequency, half-power bandwidth and centre-frequency were determined.

7.2.2.3 Feature extraction procedure

The following sequence of steps describes in detail the processing that was performed to estimate characteristic features from each sound record. This processing was implemented in the SIGPROC signal processing language (§1.3.4). Each 0.25 second long sound record, which I denote here by $s(t)$, was processed in turn by this procedure, with the resulting feature variables being stored for later statistical analysis (§7.2.2.5).

1. The RMS envelope $R(t)$ is computed by means of the technique described in §3.1.1. Each sample in $R(t)$ corresponds to the RMS value of a 50μs segment of $s(t)$ after the data values within the segment have been multiplied by a Hanning window. $R(t)$ is sub-sampled by a factor of 8 times relative to $s(t)$.

2. The start of the k^{th} click, $k = 1 \ldots N_c$, where N_c is the number of clicks in the sound record, is located at the first instant $T_{cs}^{(k)}$ at which

$$R(t) > 1/3 \max\{R(t)\}, \quad t > T_{cs}^{(k-1)} + \tau_{cl} \tag{7.3}$$

where $\tau_{cl} = 0.8$ms is the duration of each extracted click. The k^{th} click envelope $r_c^{(k)}(t)$ is defined as

$$r_c^{(k)}(t) = R(t - (T_{cs}^{(k)} - \tau_{\text{ofs}})), \quad 0 \leq t \leq \tau_{cl} \tag{7.4}$$

where $\tau_{\text{ofs}} = 50\mu$s is an offset to account for the finite rise-time of the click envelope.

3. Click intervals $I_c^{(k)}$ are defined by

$$I_c^{(k)} = T_{cs}^{(k+1)} - T_{cs}^{(k)}, \quad k = 1 \ldots N_c - 1 \tag{7.5}$$

and the average $\overline{I}_c$, standard deviation σ_{I_c}, maximum $I_{c\max}$, and minimum $I_{c\min}$ are evaluated. The average trend in the click intervals is defined by

$$\overline{dI}_c = \langle I_c^{(k+1)} - I_c^{(k)} \rangle_{k=1\ldots N_c-2} \tag{7.6}$$

Note that if $N_c = 1$, the click interval is undefined, so each of the variables defined in this step are set to zero. Likewise if $N_c = 2$, $\overline{dI}_c$ is set to zero.

4. The average click envelope

$$R_c(t) = \langle r_c^{(k)}(t) \rangle_k \tag{7.7}$$

is computed and the number of peaks N_{pc} in $R_c(t)$ of amplitude greater than $1/3 \max\{R_c(t)\}$ obtained. For each peak $j = 1 \ldots \min(4, N_{pc})$, ordered according to their amplitudes, the peak amplitude A_{pcj}, position T_{pcj}, and decay-time D_{pcj} are estimated. The decay time D_{pcj} is defined as the time taken for the envelope to decay to a value of $0.4A_{pcj}$. If $N_{pc} < 4$, the remaining variables (A_{pcj}, T_{pcj}, and D_{pcj}, $j = N_{pc} \ldots 4$) are set to zero.

5. Treating $R_c(t)$ as a probability density function (pdf) means that an equivalent time duration Δt for the click can be estimated by considering the variance of the signal about its mean (Gabor, 1946). Δt is defined as $\Delta t = \sigma_t / 2\pi$, where σ_t is the standard deviation of the normalised click envelope $\tilde{R}_c(t)$, defined by

$$\tilde{R}_c(t) = \hat{R}_c(t) / \int_0^{0.4ms} \hat{R}_c(t) \mathrm{d}t \tag{7.8}$$

where $\hat{R}_c(t)$ is the click envelope, modified according to

$$\hat{R}_c(t) = \begin{cases} R_c(t), & 0 < t < 4\text{ms}, R_c(t) > 1/10 \max\{R_c(t)\} \\ 0, & \text{otherwise} \end{cases} \tag{7.9}$$

The envelope is modified in this manner so that σ_t is not overly increased by noise in the signal. This admittedly crude and *ad hoc* modification was found to be necessary because the duration, 0.8ms, of $R_c(t)$ is much longer than the duration of most of the clicks, so that noise near the end of $R_c(t)$ dramatically increases σ_t. See §7.2.3.1 for more discussion of this point.

6. The power spectrum $P^{(k)}(f)$ of the k^{th} click is defined by the squared magnitude of the Fourier transform of a 0.4ms segment of sound centred at the position T_{pc1} of the largest peak in the envelope of the k^{th} click:

$$P^{(k)}(f) = |\mathcal{F}\{s(t) w_f(t - (T_{cs}^{(k)} - \tau_{\text{ofs}} + T_{pc1}))\}|^2 \tag{7.10}$$

where $w_f(t)$ is a 3-term Blackman-Harris window of duration 0.4ms. The windowed segment is zero-extended so that it consists of a total of 4096 samples before computing $P^{(k)}(f)$ by means of the FFT algorithm.

7. The peak frequency $f_p^{(k)}$ is obtained from each click spectrum $P^{(k)}(f)$. The average $\overline{f}_p$, standard deviation σ_{f_p}, and average trend $\overline{df}_p$ of the ensemble of $f_p^{(k)}$ estimates are computed in a similar way to that described in step 3 for the click interval variables.

8. The dominant frequency of each click is also characterised by estimating the period of oscillation of the click waveform. The zero-crossing rate $f_{zc}^{(k)}$ is defined as half the reciprocal of the average interval between zero-crossings in the click

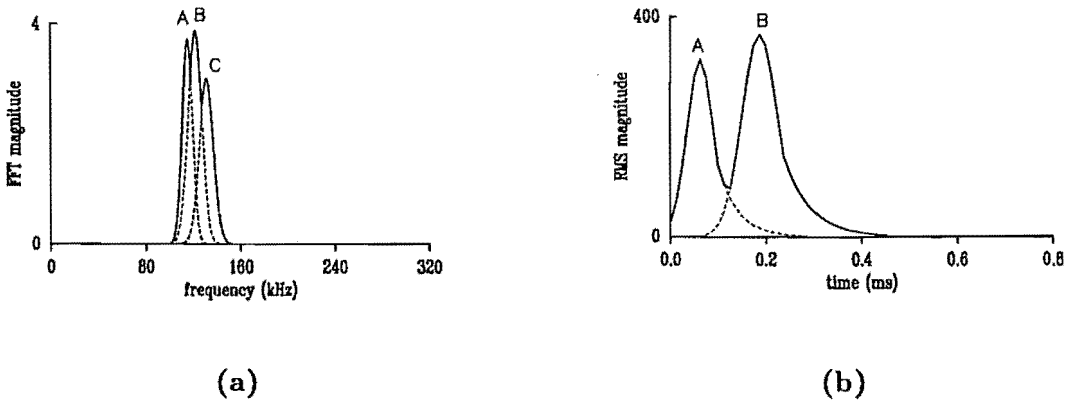


Figure 7.9. Reconstruction of click envelopes and spectra from the feature variables. **a:** Spectral magnitude, generated by overlaying a Gaussian-shaped pulse for each of the spectral peaks. **b:** Time domain envelope, reconstructed by overlaying a Gaussian pulse with exponential decay for each of the peak estimates.

waveform $s_c^{(k)}(t)$ during the interval for which $r_c^{(k)}(t) > 1/3 \max\{r_c^{(k)}(t)\}$, where $s_c^{(k)}(t)$ is the sound waveform corresponding to the click interval defined in (7.4). The average $\overline{f}_{zc}$, standard deviation $\sigma_{f_{zc}}$, and average trend $\overline{df}_{zc}$ of the N_c values of f_{zc} are computed as described in step 3.

9. The average power spectrum $P(f)$ of the sound record is obtained by averaging the $P^{(k)}(f)$, for $f > 0$. The number of peaks with amplitude greater than $1/3 \max\{P(f)\}$ is defined as N_{pf}. For the peaks of greatest amplitude $j = 1 \ldots \min(4, N_{pf})$, the amplitude A_{pfj}, frequency f_{pj}, half-power bandwidth B_{pfj}, and centre frequency f_{cj} are estimated. f_{cj} is defined as the frequency mid way between the half-power frequencies either side of f_{pj}. The centre frequency provides an indication of any assymetry in the spectral peak. As in step 4, if $N_{pf} < 4$, the remaining variables (labelled by $j = N_{pf} \ldots 4$) are set to zero.

10. The average spectrum $P(f)$ can be treated as a pdf $\tilde{P}(f)$ by defining

$$\tilde{P}(f) = P(f) / \int_0^\infty P(f) \mathrm{d}f. \tag{7.11}$$

Standard statistical formulae can be invoked to obtain the mean frequency $\overline{f}_{\mathrm{psd}}$ and standard deviation $\sigma_{f_{\mathrm{psd}}}$ of $\tilde{P}(f)$. The equivalent bandwidth of $P(f)$ is then defined as $\Delta f = 2\sigma_{f_{\mathrm{psd}}}$ (Gabor, 1946).

Table 7.1 itemises all the feature variables invoked in this section. It also lists the variable values for the sounds illustrated in Figs.7.4 and 7.5.

7.2.2.4 Reconstructing the clicks from the feature variables

In order to discover how well the extracted features characterise the clicks, synthetic versions (shown in Fig.7.10 and 7.11) of the signals shown in Fig.7.4 and 7.5 respectively were reconstructed from the features.

The spectral magnitude is reconstructed from the frequencies, amplitudes and half-power width estimates by overlaying a Gaussian shaped pulse, of the appropriate

Variable	Description	Examples		units
		Fig.7.4	Fig.7.5	
N_c	Number of clicks in sound record	3	8	
$\overline{I}_c$	Mean inter-click interval	103	30.4	ms
σ_{I_c}	Inter-click interval standard deviation	7.93	1.27	ms
$I_{c\min}$	Minimum inter-click interval	95.2	28.9	ms
$I_{c_{\max}}$	Maximum inter-click interval	111	33.1	ms
$\overline{dI}_c$	Mean trend of inter-click interval	15.9	0.7	ms/click
N_{pc}	Number of major peaks in average RMS envelope	1	2	
$A_{pc1}\ldots A_{pc4}$	Amplitudes of 4 largest peaks	367,0,0,0	368,323,0,0	
$T_{pc1}\ldots T_{pc4}$	Positions of 4 largest peaks	87.5,0,0,0	187.5,62.5,0,0	μs
$D_{pc1}\ldots D_{pc4}$	Decay-times of 4 largest peaks	37.5,0,0,0	50.0,37.5,0,0	μs
dT_{12}, dT_{13}, dT_{23}	Differences between positions of 3 largest peaks	–,–,–	125,–,–	μs
σ_{f_p}	Standard deviation of peak frequencies in record	0.99	3.0	kHz
$\overline{f}_p$	Mean of peak frequencies	125	118	kHz
$\overline{df}_p$	Mean trend of peak frequencies	-0.86	-0.80	kHz
$\sigma_{f_{zc}}$	Standard deviation of zero crossing rates of clicks in record	1.13	2.58	kHz
$\overline{f}_{zc}$	Mean of zero crossing rates	124	120	kHz
$\overline{df}_{zc}$	Mean trend of zero crossing rates	-1.2	-0.77	kHz
N_{pf}	Number of peaks in average spectrum	1	3	
$A_{pf1}\ldots A_{pf4}$	Amplitudes of 4 largest peaks	20.2,0,0,0	15,13.9,9.1,0	
$f_{p1}\ldots f_{p4}$	Frequencies of 4 largest peaks	125,0,0,0	121,115,131,0	kHz
$B_{pf1}\ldots B_{pf4}$	Bandwidths of 4 largest peaks	14.9,0,0,0	14.8,22.3,25.9,0	kHz
$f_{c1}\ldots f_{c4}$	Centre frequencies of 4 largest peaks	124.6,0,0,0	119,122,123,0	kHz
df_{12}, df_{13}, df_{23}	Differences between frequencies of 3 largest peaks	–,–,–	-6.6,9.5,16.1	kHz
Δf	Equivalent bandwidth of average spectrum	15.2	32.5	kHz
$\overline{f}_{\text{psd}}$	Mean frequency of average spectrum	125	121	kHz
Δt	Equivalent click duration of average RMS envelope	1.4	1.57	ms
$\Delta t \Delta f$	Time–bandwidth product	21.2	51	

Table 7.1. Feature variable values characterising the dolphin clicks shown in Figs.7.3 and 7.5.

height and width, for each of the peaks $j = 1\ldots4$. Each Gaussian peak $|S_{Gj}(f)|$ is described by

$$|S_{Gj}(f)| = \frac{1}{g_{fj}\sqrt{2\pi}} e^{-\frac{(f-f_{pj})^2}{2g_{fj}^2}} \tag{7.12}$$

where f_p is the peak frequency of the pulse and g_{fj} is related to the bandwidth B_{pfj}

of the j^{th} peak by the expression

$$g_{fj} = \frac{B_{pfj}}{2\sqrt{2\ln(0.5)}}. \tag{7.13}$$

Note that (7.13) defines the standard deviation g_{fj} of the gaussian peak such that its half power bandwidth is equal to B_{pfj}. The maximum amplitude of $|S_{Gj}(f)|$ is normalised to the estimated amplitude of the spectral peak $\sqrt{A_{pfj}}$. The total total reconstructed spectral magnitude $|S_{G_T}(f)|$ is given by

$$|S_{G_T}(f)| = \max_j |S_{Gj}(f)|. \tag{7.14}$$

The time domain envelope is reconstructed similarly by overlaying a pulse, consisting of a Gaussian combined with an exponential decay, for each of the component peaks of the click (Fig.7.9). Each Gaussian pulse is described by expressions similar to (7.12) and (7.13), with f, $S_{Gj}(f)$, g_{fj}, f_{pj}, and B_{pfj} replaced by their time domain equivalents t, $s_{Gj}(t)$, g_{tj}, T_{pcj}, and D_{pcj} respectively (see §7.2.2.3 and Table 7.1 for a description of each of these variables). The exponential decay for each peak is described by

$$s_{ej}(t) = e^{-D_{pcj}(t-T_{pcj})}, \quad t > T_{pcj}. \tag{7.15}$$

The total time domain envelope is constructed by overlaying $s_{Gj}(t)$ and $s_{ej}(t)$, for $j = 1 \ldots 4$, in a similar way as described by (7.14) for the frequency domain data. An exponential decay was employed in the time domain envelope reconstruction because it fitted with the observed data and was simple to calculate. In addition, it accords with the decay of oscillations caused by a damped resonating sound generator (or equivalently by decaying echoes within the dolphin's head — see §7.1.4.1).

Figs.7.10 and 7.11 show the reconstructed time and frequency domain envelopes corresponding to the feature variables obtained from the clicks shown in Figs.7.4 and 7.5 respectively. The similarities between the original and reconstructed versions of these clicks indicate the ability of the feature variables invoked here to to describe the shapes of the click envelopes.

7.2.2.5 Comments on the automatic extraction of descriptive features

Automatic implementation of the feature estimation process requires the programs to be capable of processing the entire range of sound records. The programs were written in a modular fashion, and the processed signals (prior to the actual extraction of the feature variables) were saved on magnetic tape. If any part of the extraction procedure was later found to require modification, it was a simple matter to repeat that procedure on the processed signals.

After each set of features was extracted from a sound record, it was written out as a text file, together with a label denoting the original tape and tape counter number corresponding to the sound record (see §7.2.1). These labels enabled the matrix of feature variables to be easily integrated into an already existing database (on a Macintosh SE micro-computer) that listed the behavioural contexts of each recording.

The labels also facilitated the management of the data. This was important due to the large amount of data (about 430 raw data records of 300 kBytes each). The processing was performed in batches of about 60 records each, with the rest of the data stored on magnetic tape. To avoid having to re-process the raw data if any of the analysis parameters required adjustment, the extracted clicks, average time envelopes, average spectral magnitudes and records of inter-click intervals for all clicks were stored separately from the raw data records. By means of this data compression (effectively

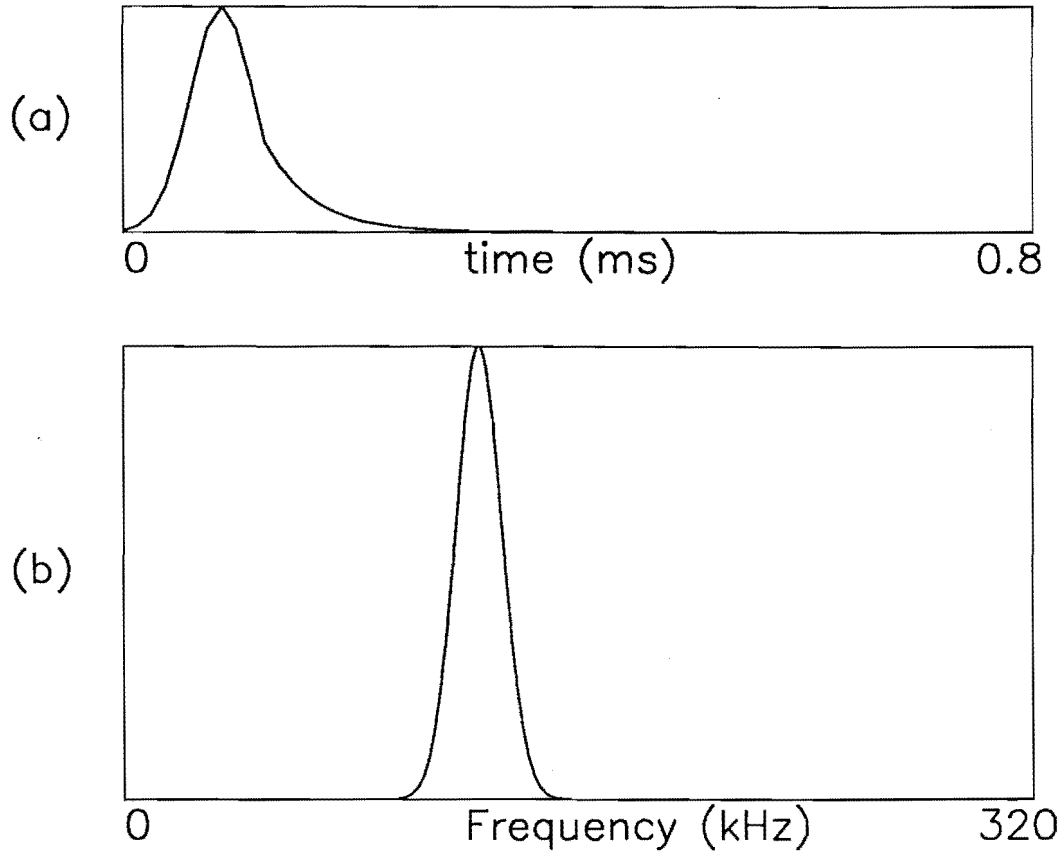


Figure 7.10. Reconstruction of the click shown in Fig.7.4. **a:** Average envelope profile and **b:** average spectrum.

removing the gaps between the clicks) the processed data for all records could be stored in about 20 MBytes.

In order to check the integrity of the records employed in the statistical analysis, stackplots of the click waveforms, and the average click envelope spectrum from each sound record were visually inspected on a graphics work-station. This identified any records containing overlapping clicks from several dolphins, or those that were corrupted by unduly high levels of noise. The records so identified were discarded from the data set.

7.2.3 Statistical analysis of Hector's dolphin acoustic repertoire

7.2.3.1 Descriptive statistics of sound features

The 401 records that were analysed contained 7661 "clicks", with the average interval between clicks in a record ranging from 1.3ms to 164ms (median = 27.6ms). The average frequency of the clicks ranged from 82kHz to 135kHz, with a mean of 125kHz. Most of the records had clicks with one (52%) or two (36%) peaks in their energy envelope, and 92% had one or two peaks in their spectrum (Dawson and Thorpe, 1990).

Histograms of the click time duration and bandwidths are shown in Fig.7.12 and 7.13. The histograms illustrate that most of the clicks are short and narrow-band. The few that are very wide-band and/or of long duration generally seem to be "noisy" signals. One of the very wide-band signals contained a strong low frequency component. This is described in §7.2.4. Noise served to increase the standard-deviation

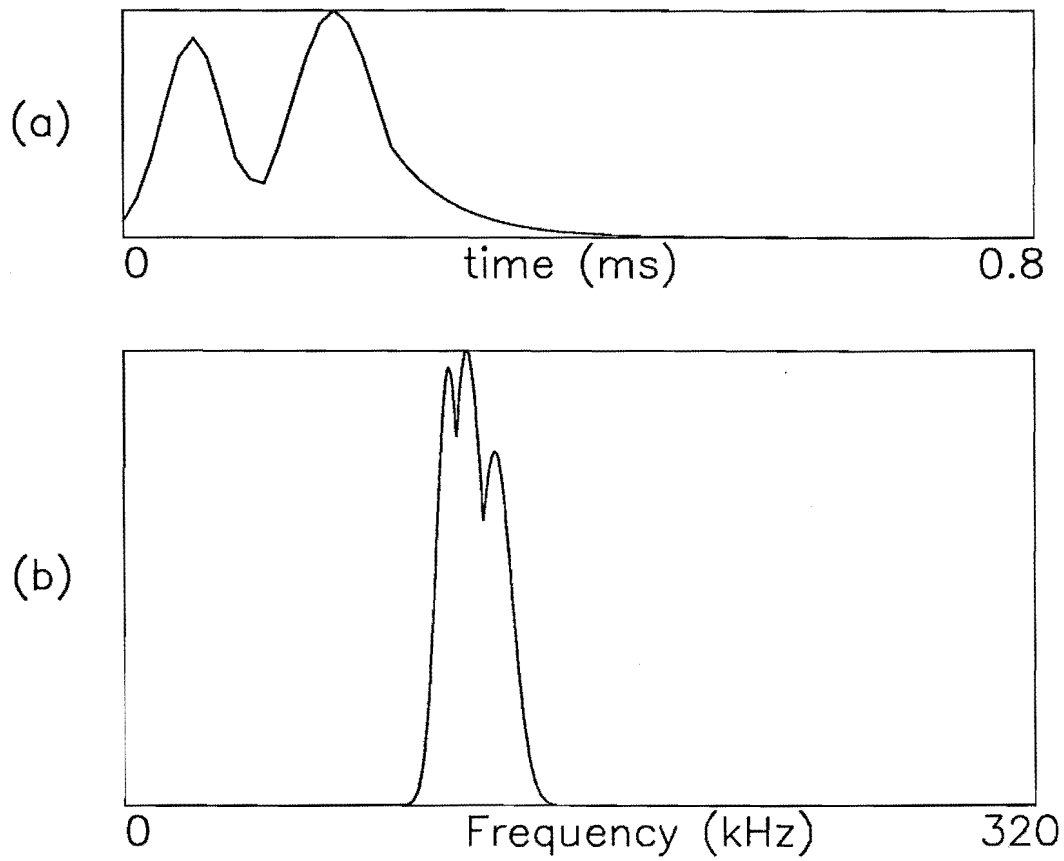


Figure 7.11. Reconstruction of the click shown in Fig.7.5. **a:** Average envelope profile and **b:** average spectrum.

in both the time and frequency domains. The effect of noise and echoes on the time duration feature was reduced by setting everything in the average signal envelope that was below a certain threshold to zero before calculating the variation.

The time–bandwidth product gives an indication of how "complicated" a signal is. Gabor (1946) showed that there is a minimum value that indicates the "simplest" type of signal. Because of the way in which it is defined here (§7.2.2), the minimum value is unity. A histogram of the time–bandwidth values obtained for the clicks analysed here is shown in Fig.7.14. While most of the signals have a near-minimum time–bandwidth product, some values are considerably larger. The values are much greater than those obtained by Wiersma (1982,1988) for various other odontocete sonar signals, which were all between 1.1 and 1.5. This seems to suggest that Hector's dolphin sonar signals are more "complicated" than the ones that Wiersma studied. However, as discussed in the previous paragraph, the ones with very large values generally consisted of noisy signals.

Part of the difference between the time–bandwidth values described by Wiersma and those presented here could arise from the differences between our respective measuring techniques. Wiersma manually selects a segment of each click that encompasses the main "lobe" of the click but excludes the subsequent, lower amplitude, "decay" of the waveform (see Fig.7.5*b*). The lower amplitude parts of the waveform are assumed to arise from echoes within the dolphin's head (§7.1.4.1). By using this technique, it is not suprising that he obtains near-minimum values for the time–bandwidth product. I think that the time–bandwidth product should include the entire signal as emitted

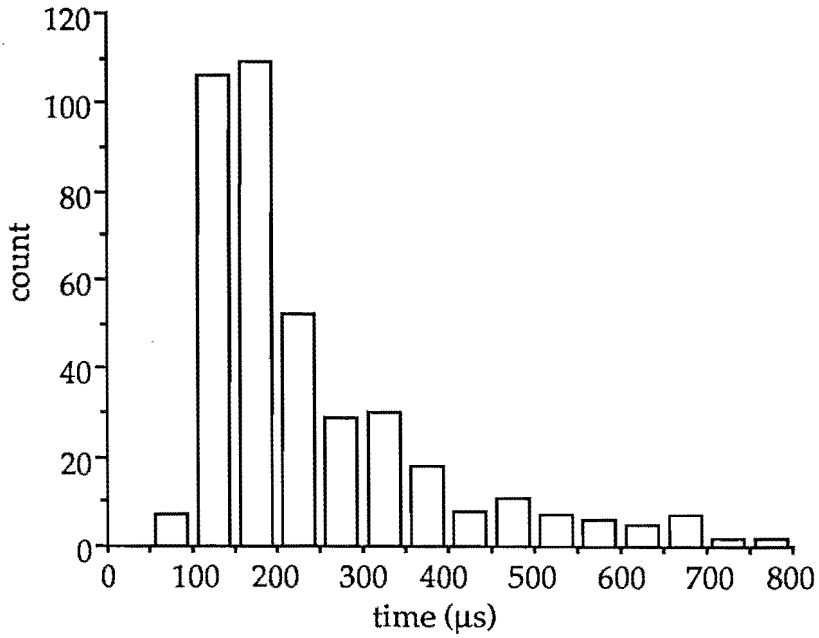


Figure 7.12. Histogram of the time durations Δt of the dolphin clicks.

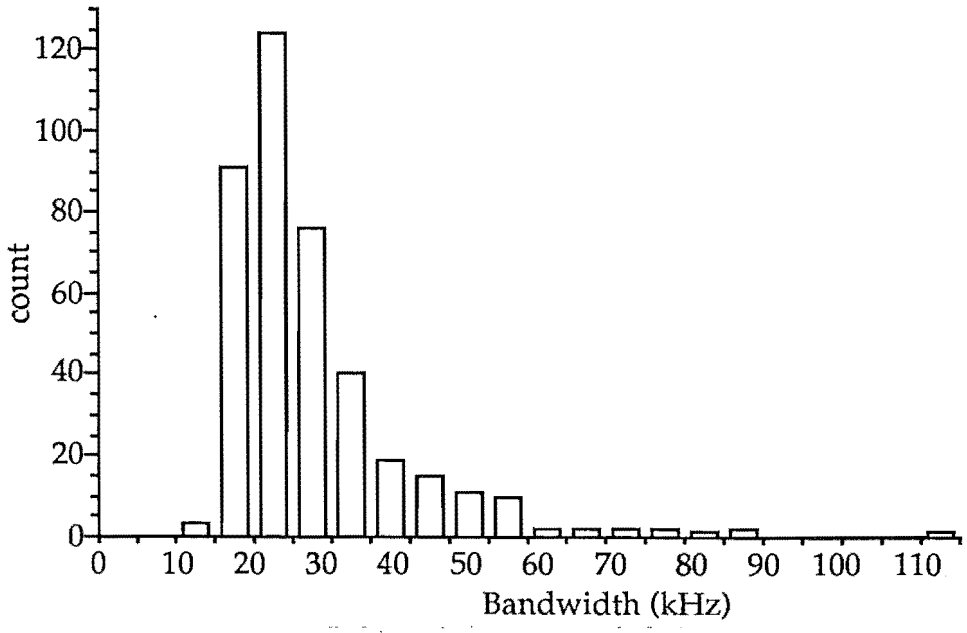


Figure 7.13. Histogram of the spectral bandwidths Δf of the dolphin clicks.

by the dolphin, since this is what the dolphin actually uses for its echo-location (see also Kamminga and Wiersma, 1981). However, the few very large values of Δf and Δt that I obtained imply that the statistical approach to measuring bandwidth and time duration is not appropriate when the signal consists of several components or peaks, or is corrupted with significant amounts of noise. Because of my automatic approach to evaluating the feature variables, I employed a threshold to exclude noise in the signal (§7.2.2.3). A simple automatic approach such as this cannot hope to be as accurate as manual editing of each individual click waveform.

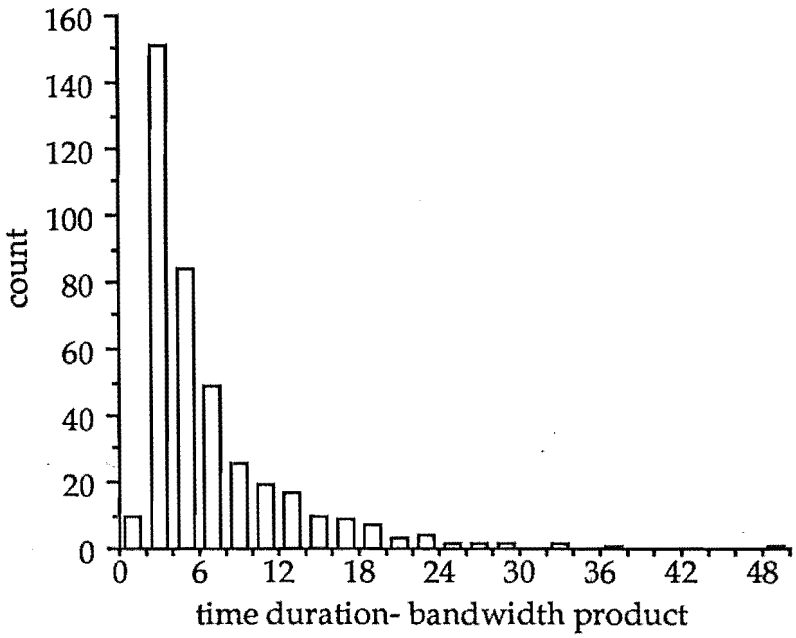


Figure 7.14. Histogram of the time–bandwidth products $\Delta t \Delta f$ of the clicks.

7.2.3.2 Principal Component Analysis

Principal component analysis (PCA) is a method of reducing the dimensionality of a feature set by determining the orthogonal "principal components" which describe most of the variance in the data set (Cooley and Lohnes, 1971). Typically only the largest two or three principal components (often termed *factors*) are retained. The feature vectors can be projected onto the principal component axes and plotted as points in a two or "three" dimensional space, in order to indicate the occurrence of clusters or patterns in the data (§1.3.3). The relationship between the original feature vectors and each of the principal components is indicated by the *factor loadings*, which are simply the factors expressed as vectors in the original feature space.

The magnitude of the factor loading for a particular feature variable indicates the degree to which that feature is correlated with the factor. Comrey (1973) suggests that the magnitude of a factor loading can be interpreted as representing "good", "very good", or "excellent" correlation between a variable and a factor if it is greater than 0.55, 0.63, or 0.71 respectively. In the PCA results presented here, variables are associated with the factors for which their factor loadings are greater then 0.55.

In general, it is difficult to interpret the physical significance of the principal components identified by PCA, because each one contains contributions from all the physical feature variables. In order to facilitate the interpretation of the PCA, the axes can be rotated so that each one is, as much as is possible, composed of a different subset of the feature variables. The "varimax"' rotation (Cooley and Lohnes, 1971, §5.3) was invoked in this analysis. Each of the resulting rotated factors can then be associated with a separate physical interpretation.

Principal component analyses were conducted using the Systat statistics package (Wilkinson, 1987) running on a Macintosh SE micro-computer. Two PCAs were performed, one with variables describing the time domain characteristics of the clicks, and the other with variables that described the frequency domain characteristics.

Table 7.2 shows the loadings of the time domain variables for each of the first three factors. Together these three factors account for 74.6% of the total variance in the set of data. The loadings indicate that factor 1 represents the amplitude and

variable	Factor 1	Factor 2	Factor 3
T_{pc1}	-0.073	-0.041	0.456
D_{pc1}	0.129	0.065	**0.725**
A_{pc2}	0.240	**0.723**	0.380
T_{pc2}	0.110	**0.977**	-0.015
D_{pc2}	0.305	**0.642**	0.458
A_{pc3}	**0.775**	0.171	0.425
T_{pc3}	**0.965**	0.191	0.056
D_{pc3}	-0.001	0.107	**0.703**
dT_{12}	0.057	**0.932**	-0.168
dT_{13}	**0.960**	0.185	-0.060
dT_{23}	**0.936**	0.040	-0.135
Variance explained	32%	26%	16.5%

Table 7.2. Rotated PCA factor loadings for the time domain analysis. High correlations between variables and factors are identified by bold type.

variable	Factor 1	Factor 2	Factor 3	Factor 4
f_{p1}	-0.260	-0.186	**-0.878**	-0.215
A_{pf1}	0.014	-0.011	-0.136	**-0.710**
B_{pf1}	0.220	0.199	-0.056	**0.791**
f_{c1}	-0.170	0.040	**-0.960**	0.052
f_{p2}	0.043	**0.962**	0.106	0.152
A_{pf2}	0.014	**0.808**	-0.096	-0.225
B_{pf2}	0.218	**0.766**	0.140	0.441
f_{c2}	0.068	**0.958**	0.114	0.158
f_{p3}	**0.969**	0.075	0.170	0.072
A_{pf3}	**0.852**	0.125	-0.092	-0.125
B_{pf3}	**0.862**	0.034	0.227	0.306
f_{c3}	**0.967**	0.074	0.177	0.078
df_{12}	0.048	**0.644**	**0.690**	0.093
df_{13}	**0.657**	0.008	**0.683**	0.052
df_{23}	**0.854**	0.025	0.406	0.168
Variance explained	31%	24%	20%	11%

Table 7.3. Rotated PCA factor loadings for the frequency domain analysis. High correlations between variables and factors are identified by bold type.

position of the third peak relative to the first and second peaks, factor 2 represents the characteristics of the second peak, while factor 3 represents the decay (width) of the first and third peaks. A scatter plot of the first two factors (Fig.7.15) indicates three broad groups of sounds, corresponding to those with 1, 2, or 3 peaks respectively in their time domain amplitude envelopes. The sounds are spread out along the factor 2 axis according to the separation of the first two peaks, and along the factor 1 axis according to the distance of the third peak (if present) from the first two peaks. The dense cluster of points in the lower left of the figure represents sounds that only have

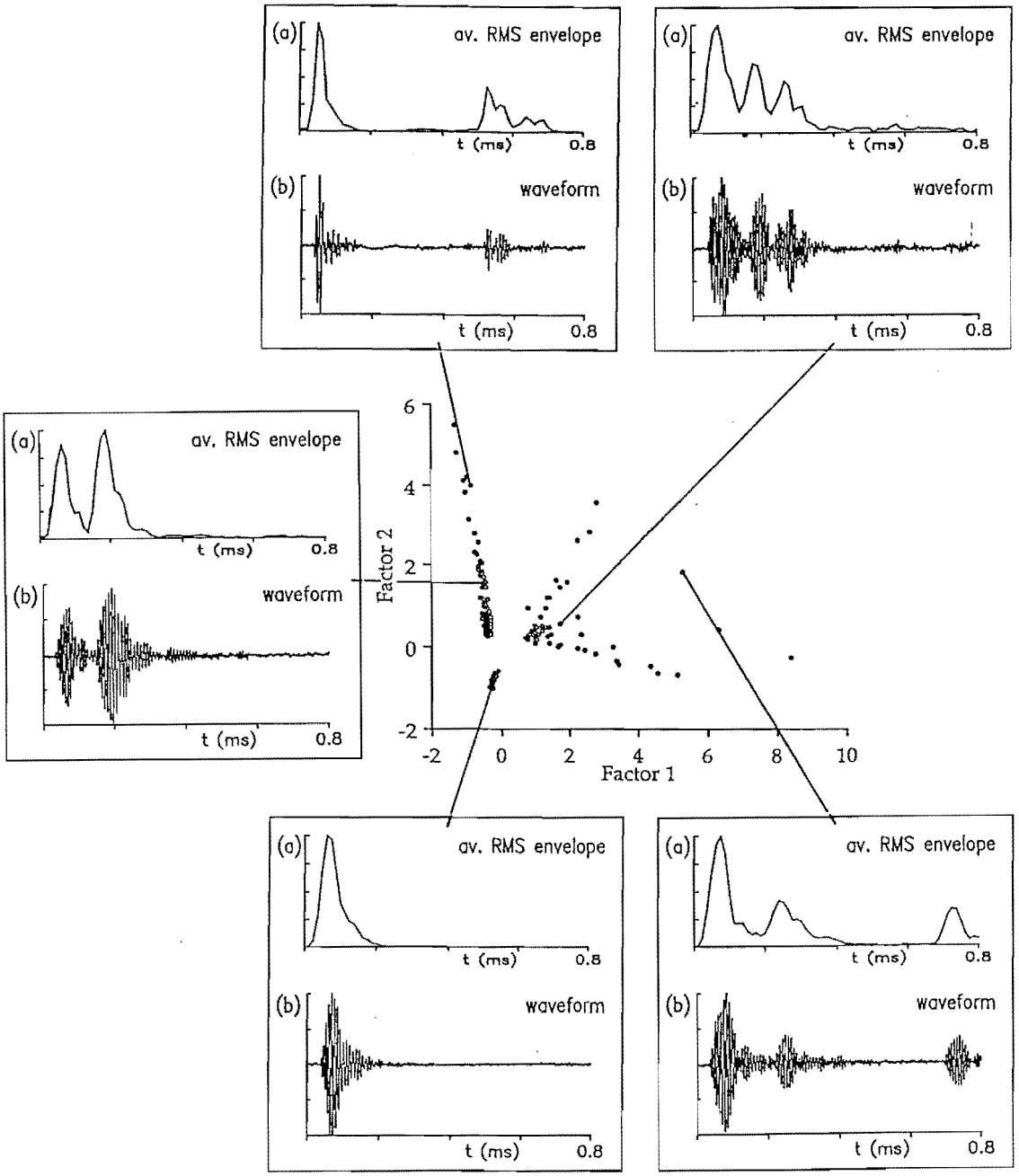


Figure 7.15. Graph of the data records relative to the first two factors of the PCA conducted on the variables obtained from the average envelope of each record. Example waveforms and energy envelopes are also shown, together with the position of each example in the scatter plot.

a single peak.

PCA of the variables that describe the shape of the average click spectrum of each record revealed four significant factors, that together accounted for 86% of the total variance in the data set (after rotation). The factor loadings (Table 7.3) indicate that factor 1 represents the characteristics of the third spectral peak, if it is present, while factor 2 represents the characteristics of the second peak. Factor 3 represents the peak and centre frequencies of the largest peak, and factor 4 represents the amplitude and half-power width of the largest peak.

Fig.7.16 shows a scatter plot of the sound records according to their first two

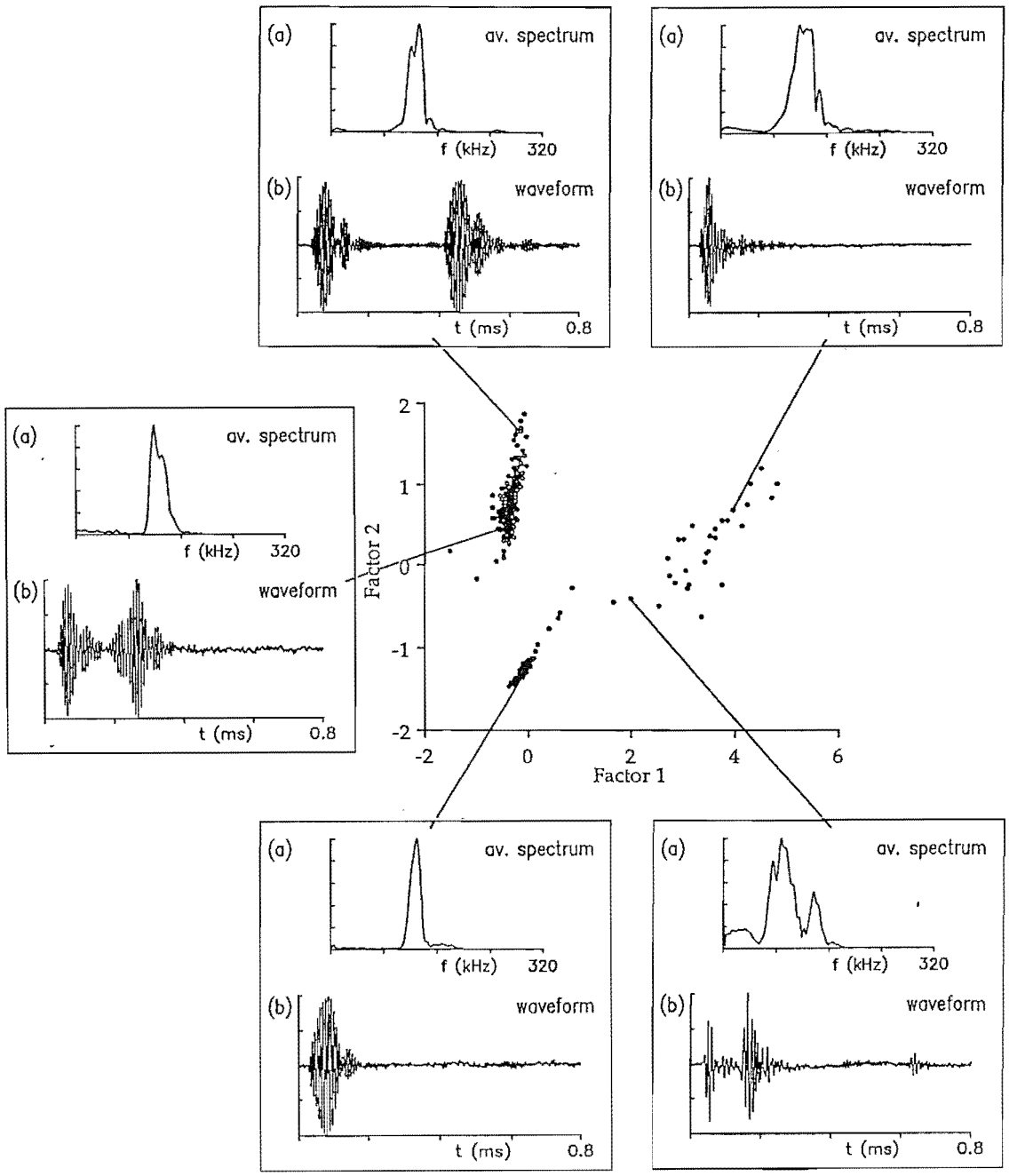


Figure 7.16. Graph of the data records relative to the first two factors of the PCA conducted on the variables that describe the shape of the average power spectrum of each record. Example waveforms and spectra are also shown, together with the position of each example in the scatter plot.

frequency domain PCA factors. This indicates that the sounds are largely characterised by their number of spectral peaks. Sounds whose spectra consist of a single peak are grouped in the tight cluster at the lower left of the scatter plot. Above this cluster is a more diffuse group of points which represents signals for which two spectral peaks were detected. Sounds with three spectral peaks are spread out in the diffuse cloud of points on the right of the scatter plot.

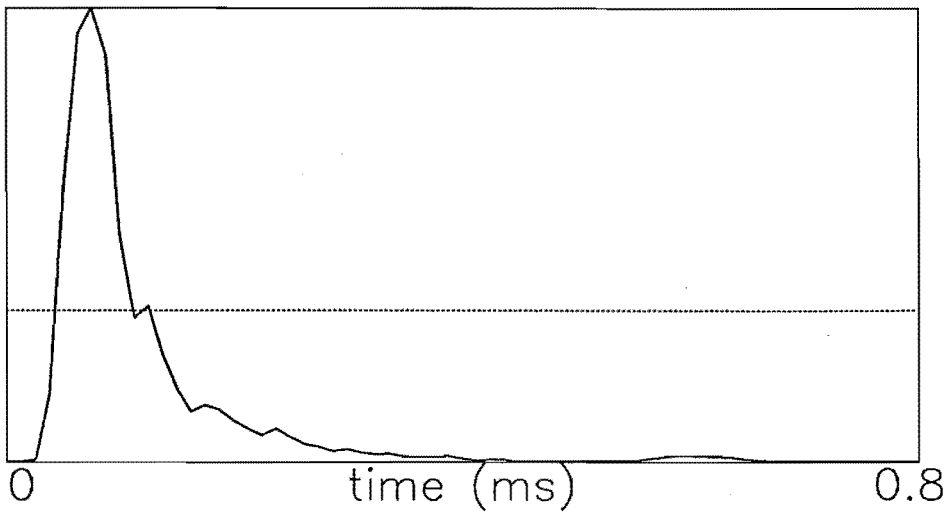


Figure 7.17. Example of a click with a single peak in its envelope that was originally classified as having two peaks. Notice the minor second peak, which is greater than 1/3rd of the major peak, so is encoded as a second feature peak.

7.2.4 Comments on repertoire characterisation

The three ways in which I measured the dominant frequency of clicks [by measuring the peak frequency for each click in a record and averaging over all the clicks in the record ($\overline{f_p}$); by measuring the mean frequency from the average spectrum of a record ($\overline{f_{\text{psd}}}$); by averaging the zero-crossing rate of each click over all clicks in a record ($\overline{f_{zc}}$)] produced results that were highly correlated ($r > 0.733$, $p < 0.001$). However, in 22 record the measurements differed by more than 10 kHz between any two of the three methods. These records often had several peaks in their spectra, and had significantly wider bandwidths Δf (mean = 52.7± 8.4 kHz [95%ci]) than the remainder (26.1 ± 0.9 kHz; t=6.57, 21df, $p < 0.001$). It is obvious that the spectral shape of such signals is not adequately represented by a simple single measurement of the dominant frequency.

One of the problems encountered in the repertoire analysis was the sensitivity of the statistical analysis to changes in the choice of features. For example, in a preliminary analysis of the time domain envelope features, there did not appear to be any clear-cut distinctions between clicks in different clusters. However, closer inspection revealed that the clusters were in fact characterised by the number of peaks in the envelope. The problem lay in the definition of "peak". Fig.7.17 shows a click that has a dominant single peak in its envelope. Because the small "bump" on the side of the main peak is slightly greater than 1/3rd the amplitude of the main peak, the feature extraction procedure of §7.2.2.3 obtained two peaks for this click. Furthermore, PCA located this click close to the one shown in Fig.7.5, rather than that of Fig.7.4 as one would expect. In order to overcome this problem, the locations of the second and subsequent peaks were replaced by their distances from the previous peaks, and the heights of each peak were specified as a proportion of the largest peak. Even though the distances between peaks are linear combinations of the actual peak positions, this re-expression appeared to "improve" the results obtained in the PCA. The PCA of the resulting data set was much more physically "satisfying" than that originally obtained, in that clicks such as the one shown in Fig.7.17 were placed closer to clicks like the one shown in Fig.7.4.

The above-mentioned problems of feature selection, and the difference in group-

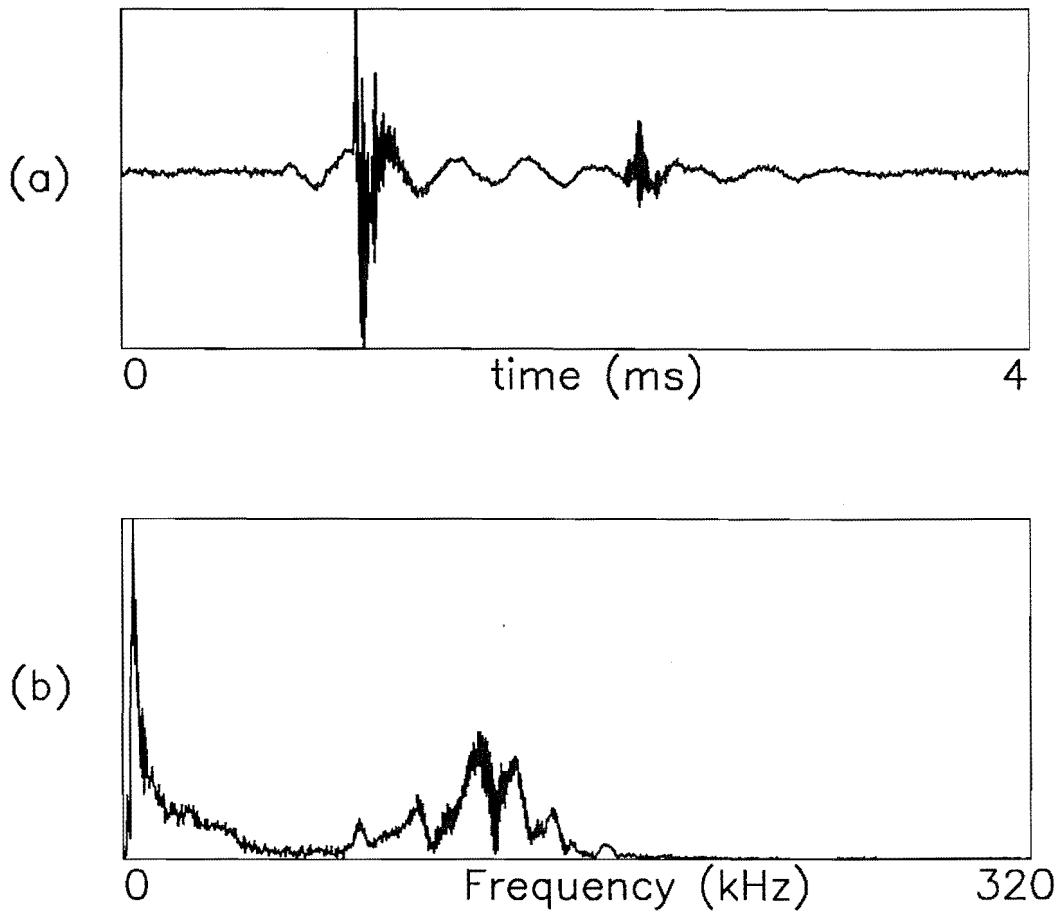


Figure 7.18. a: Example of a click with distinct high- and low-frequency components, together with b: its corresponding spectrum.

ings between the PCAs conducted on the time and the frequency domain features, illustrate the difficulties of interpretation that occur in this type of statistical analysis. The groups indicated by the statistical analysis are strongly affected by the choice of features. Because the sounds were analysed according to features that characterised the signal envelope and spectral shape, the results appear in terms of the variation of these quantities. In particular, the signal shapes were characterised by "peaks", and the classification obtained was in that form. The validity of these features as descriptive of the signal shape is confirmed by the reconstructions described in §7.2.2.4. Whether such a classification is relevant in terms of the perception of the dolphin is difficult to ascertain. For echo-location purposes, the occurrence of multiple peaks in the energy envelope is probably of importance (§7.1.4.2). The occurrence of multiple peaks in the (average) spectrum can occur either because of the presence of multiple peaks in the time domain envelope (the so-called time-separation pitch §7.1.4.2) or because of phase or frequency shifts in the time waveform. Phase shifts could be caused by multiple sources or internal reflections in the sound production mechanisms as suggested by Dziedzic (1978) and Wiersma (1982). Multiple peaks in the average spectrum can also occur if there is a difference in the peak frequencies of different clicks within a single record.

In many of the plots of statistical results, including the PCA scatter diagrams and the histograms of $\Delta t \Delta f$ and bandwidth, one click appeared on its own as an outlier.

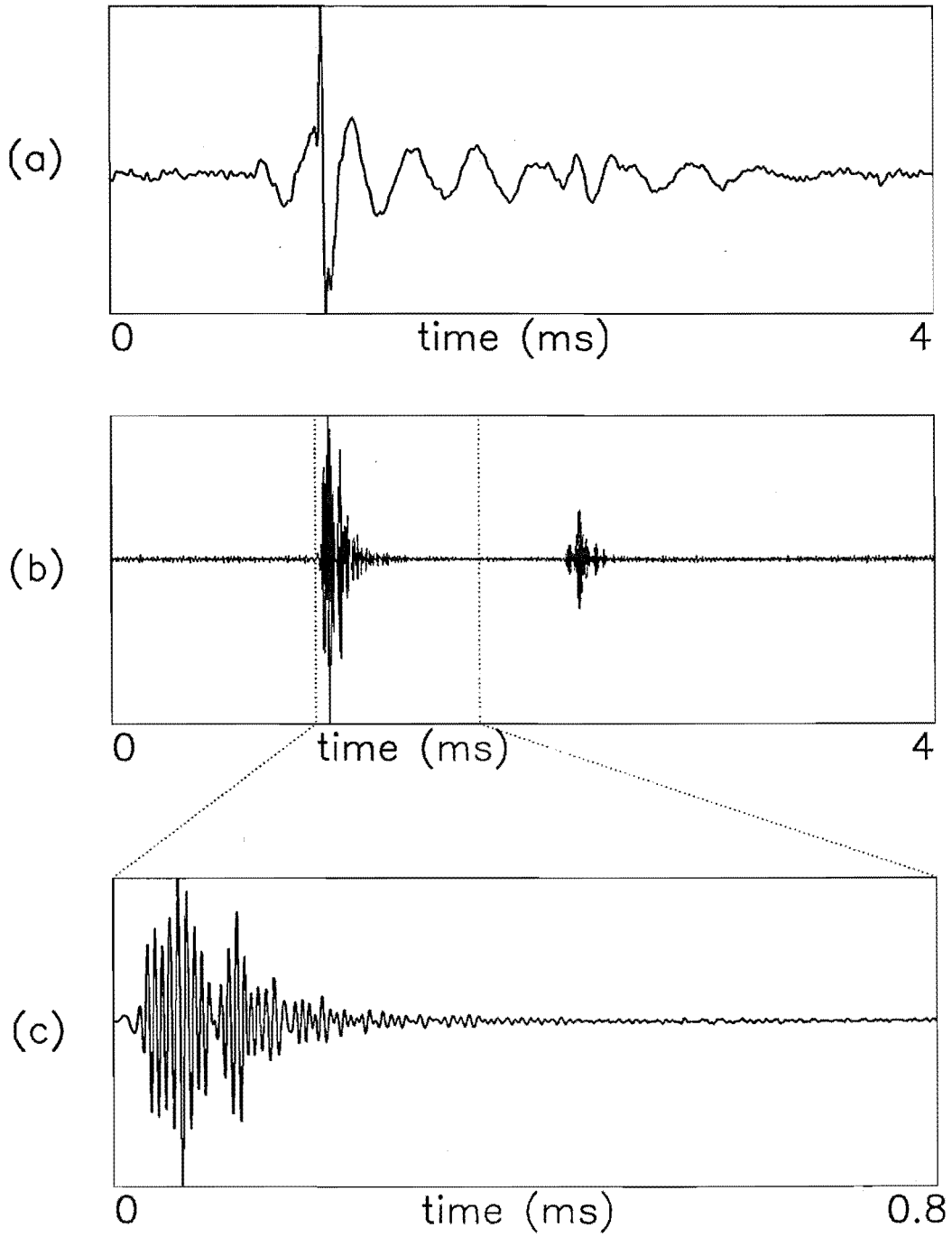


Figure 7.19. a: Low-frequency component, b: high-frequency component and c: expanded high-frequency component of the click shown in Fig.7.18.

Upon closer examination, this click was found to contain two distinct frequency components, at 5.8 kHz and 122 kHz (Fig.7.18). Separating the two components by (digital) filtering shows that the high-frequency part is similar to the other high-frequency signals that were recorded (Fig.7.19*c*). The two parts are clearly not independent, since the production of the high-frequency component severely affects the low-frequency component (Fig.7.19*a*). No other records contained low-frequency (<50 kHz) peaks greater than 1/3 of the maximum peak in the spectral magnitude.

7.3 Echo-location capability of Hector's dolphin

By treating the recorded clicks as sonar signals, the ability of Hector's dolphin to localise prey and underwater obstacles can be assessed by means of *ambiguity analysis* of the click sounds.

Unlike some other echo-locating mammals, such as a few species of bats (e.g. *Rhinolophus ferrumequinum* which employs CF-FM (constant frequency, frequency modulated) composite sonar signals, Schnitzler, 1968), and some other odontocetes (e.g. *Phocoena phocoena* which has a dual-component sonar signal, Evans, 1973), Hector's dolphins emit sounds of comparatively narrow (3dB) bandwidth (typically 15% of the frequency of the spectral peak — see §7.2.3.1). Consequently, standard narrow-band radar signal processing theory (Skolnik, 1980, Chapter 11) is sufficient for assessing the dolphin's echo-location capabilities. The narrowness of the bandwidth, and the smallness of the time–bandwidth product, of each of the recorded sounds enables the procedure outlined in §7.3.1 to be employed for this purpose. In §7.3.2 I present the results obtained by applying this technique to examples of the clicks encountered Finally, §7.3.3 attempts to elucidate aspects of the dolphin's echo-location performance from the analysis results.

7.3.1 Calculation of ambiguity surface

Calculation of the *ambiguity function* of each of the clicks proceeds according to the following series of steps:

(i) The sound is expressed in terms of its analytic signal representation $\psi(t)$ (§1.2.6):

$$\psi(t) = u(t)\, e^{(i2\pi f_o t)} \tag{7.16}$$

where f_o is the carrier frequency (defined in (iii) below) of the sound, and $u(t)$ is the waveform (i.e. the modulation on the carrier) of the sound.

(ii) $\psi(t)$ is constructed by, first, computing the spectrum $S(f)$ of the sound, then defining

$$\begin{aligned} \Psi(f) &= 0 \quad \text{for} \quad f < 0 \\ &= S(f) \quad \text{for} \quad f \geq 0, \end{aligned} \tag{7.17}$$

and finally taking $\psi(t)$ to be the inverse Fourier transform of $\Psi(f)$.

(iii) $u(t)$ is taken to be the inverse Fourier transform of $\Psi(f - f_o)$, with f_o defined as the mean value of $|\Psi(f)|$.

(iv) The narrow band ambiguity function (Woodward, 1953) is computed by application of

$$\chi(\tau, \phi) = \int_{-\infty}^{\infty} u^*(t)\, u(t+\tau) e^{(-i2\pi\phi t)}\, dt \tag{7.18}$$

where τ and ϕ represent the usual incremental delay (equivalent to target range) and doppler (equivalent to target velocity) parameters.

1. The *ambiguity density* of the sound is defined as

$$A(\tau, \phi) = |\chi(\tau, \phi)/\chi(0,0)|^2. \tag{7.19}$$

The ambiguity density sets the fundamental limit on the sound's ability to distinguish by echo-location between two targets of equal strength (i.e. equal reflectivity) separated by τ and ϕ in delay and doppler respectively (Woodward, 1953). It is

important to remember that all waveforms possess the same "total ambiguity" because

$$\int_{-\infty}^{\infty}\int A(\tau,\phi)d\tau d\phi \equiv 1 \tag{7.20}$$

The signal response function $\chi(\tau,\phi)$ was calculated by evaluating $u^*(t)u(t+\tau)$ for discrete incremental delays $\tau = \tau_n$ and then employing the FFT to generate $\chi(\tau_n,\phi)$ for each value of the integer n defined by $\tau_n = n\Delta\tau$ within the range $-\tau_{\max} < \tau_n < \tau_{\max}$, where $\tau_{\max}$ is the maximum desired temporal delay and $\Delta\tau$ is the required temporal resolution. Both $\Delta\tau$ and $\tau_{\max}$ were calculated using the usual radar range equation (§7.1.2) from the range resolution ΔR and the maximum range $R_{\max}$ respectively that were deemed to be sufficient to reveal relevant details in each ambiguity diagram. For the results reported in §7.3.2 $\Delta R = 5$mm and $R_{\max} = 200$mm.

The spacing $\Delta\phi$ of adjacent values of ϕ, at which $\chi(\tau,\phi)$ is evaluated, need be no less than the reciprocal of the effective duration of $u^*(t)u(t+\tau)$. In order to display more clearly the detail revealed in the ambiguity diagrams, and also to ensure that the number of temporal samples of $u^*(t)u(t+\tau)$ is a power of 2 (necessary to implement the FFT algorithm available to me), $u^*(t)u(t+\tau)$ was zero-extended, so that its actual duration was up to an order of magnitude longer than its effective duration. Consequently, $\Delta\phi$ was always somewhat less than both $1/T$ and the value, corresponding to a velocity resolution of 4m/s, which was deemed adequate to reveal relevant details in the ambiguity diagrams.

Note that the ambiguity diagrams displayed in Figs.7.20 to 7.22 are plotted as functions of incremental range R and velocity v, rather than delay τ and Doppler frequency ϕ. Even though target velocities exceeding 10m/s are probably of little concern to Hector's dolphin, all the detail (out to values of $|\phi|$ corresponding to velocities of 100m/s) revealed in the computed ambiguity diagrams are plotted in the figures, in order to display the complete structure of each diagram.

7.3.2 Results of ambiguity analysis

Figs.7.20*b*,*c* 7.21*b*,*c* and 7.22*b*,*c* depict the ambiguity diagrams of the clicks shown in Figs.7.20*a*, 7.21*a* and 7.22*a* respectively. Both contour plots and relief maps of the ambiguity diagrams are presented in order to display as clearly as possible the full structure of each diagram. Note that the relief maps reveal more detail. Since the reason for including the contour plots is to make it easier for the reader to interpret the relief maps, a comparatively coarse contour spacing is adequate for these. Figs.7.20*b* and *c* show an ambiguity diagram exhibiting several major peaks, which occur along the velocity and range axes, with only a little ambiguity density off these axes. The main peak has a width (measured from its centre to where it falls to half of its peak value) of 2cm along the range axis and 20m/s along the velocity axis. The other lobes all have peak amplitudes which are less than half the value of the main peak. The click responsible for this ambiguity diagram could resolve velocity differences down to 20m/s and range differences greater than 2cm. However, Hector's dolphins have a maximum swimming speed of about 10m/s (Slooten and Dawson, 1988) and so seem unlikely to have much use for such coarse velocity resolution.

The ambiguity diagram shown in Figs.7.21*b* and *c* consists almost entirely of a narrow (in range) ridge extending out to beyond 50m/s along the velocity axis, with negligible ambiguity density more than 1cm (in range) from this axis. A sonar system employing the click shown in Fig.7.21*a* would have negligible ability to resolve velocity differences and can be described as doppler insensitive. The range ambiguity of 2cm is essentially the same as that for the click shown in Fig.7.20*a*.

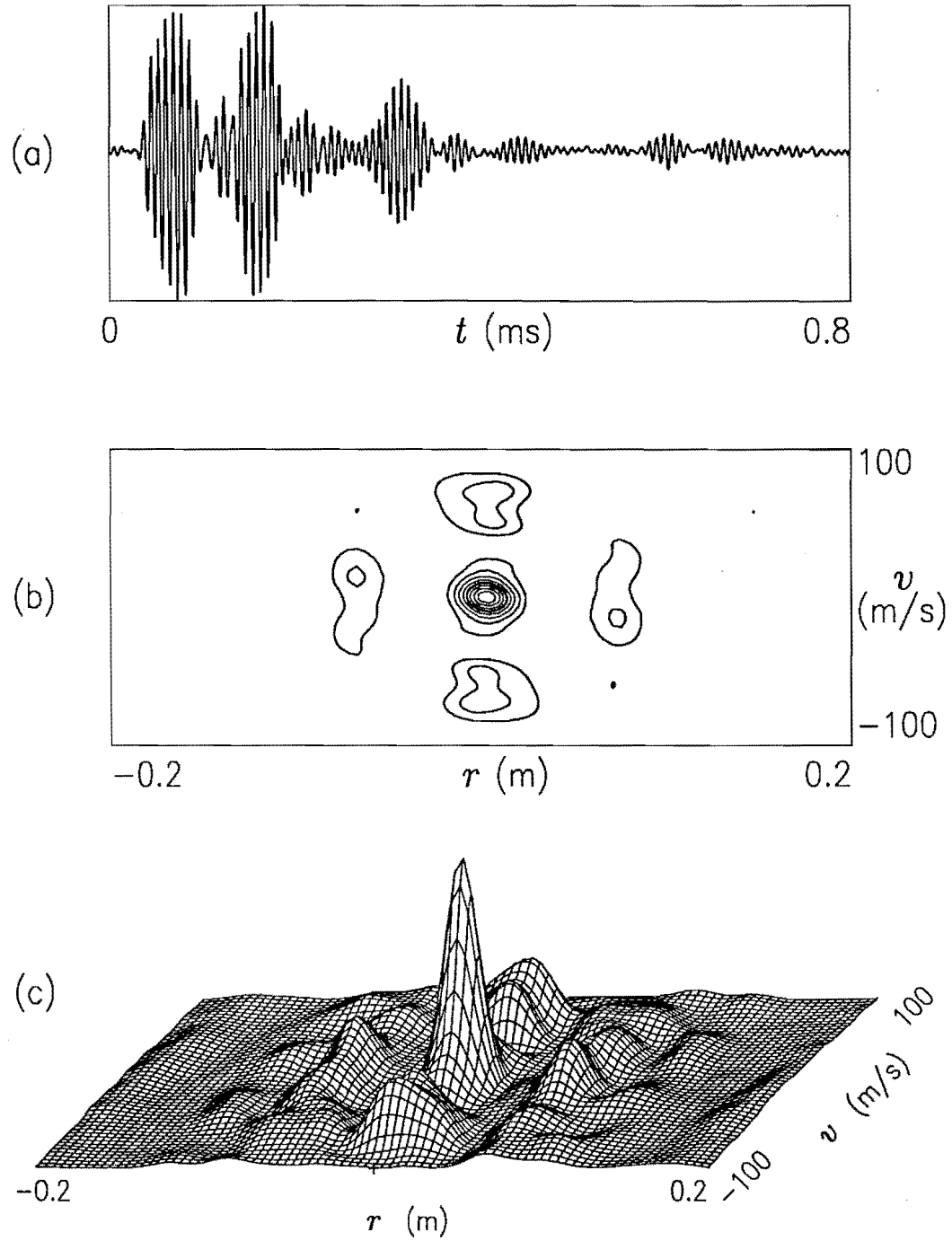


Figure 7.20. **a:** Example of a sonar click exhibiting several distinct peaks and its ambiguity diagram: **b:** contour plot (contours spaced at $A(0,0)/10$), **c:** relief map

Figs.7.22*b* and *c* show an ambiguity diagram that again consists of a single narrow ridge, but which is angled with respect to the velocity axis, due to a slight frequency sweep within the click. The extent of this sweep implies that the velocities of targets are resolvable if their differences exceed about 70m/s, which seems to be far too coarse a resolution to be of practical assistance to the dolphins in detecting real targets, whose velocities seldom exceed 10m/s. However, it may be useful in that it enables the dolphin to emit a click longer than that shown in Fig.7.21*a* (for example), thereby significantly increasing the energy in the transmitted sound, while retaining

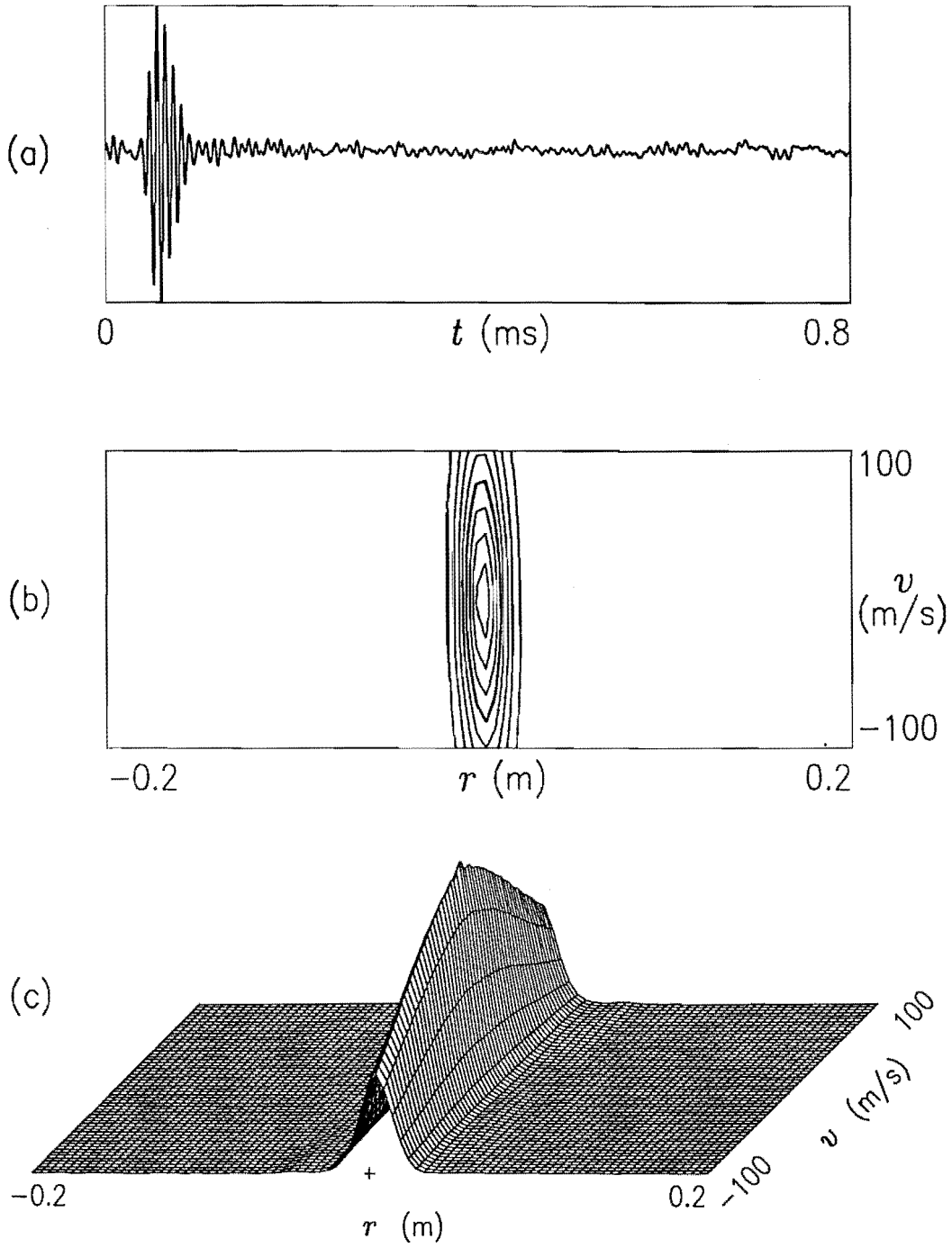


Figure 7.21. a: Example of a short sonar click and its ambiguity diagram: **b:** contour plot (contours spaced at $A(0,0)/10$), **c:** relief map

a similar range resolving ability. As indicated by the results in §7.2.3, clicks similar to that displayed in Fig.7.22*a* are much more common than clicks like those shown in Figs.7.20*a* and 7.21*a*.

7.3.3 Interpretation of ambiguity diagrams

Hector's dolphins, emitting the types of sounds examined in §7.3.2, should be capable of resolving targets down to 2cm apart. This accuracy is achieved at the expense

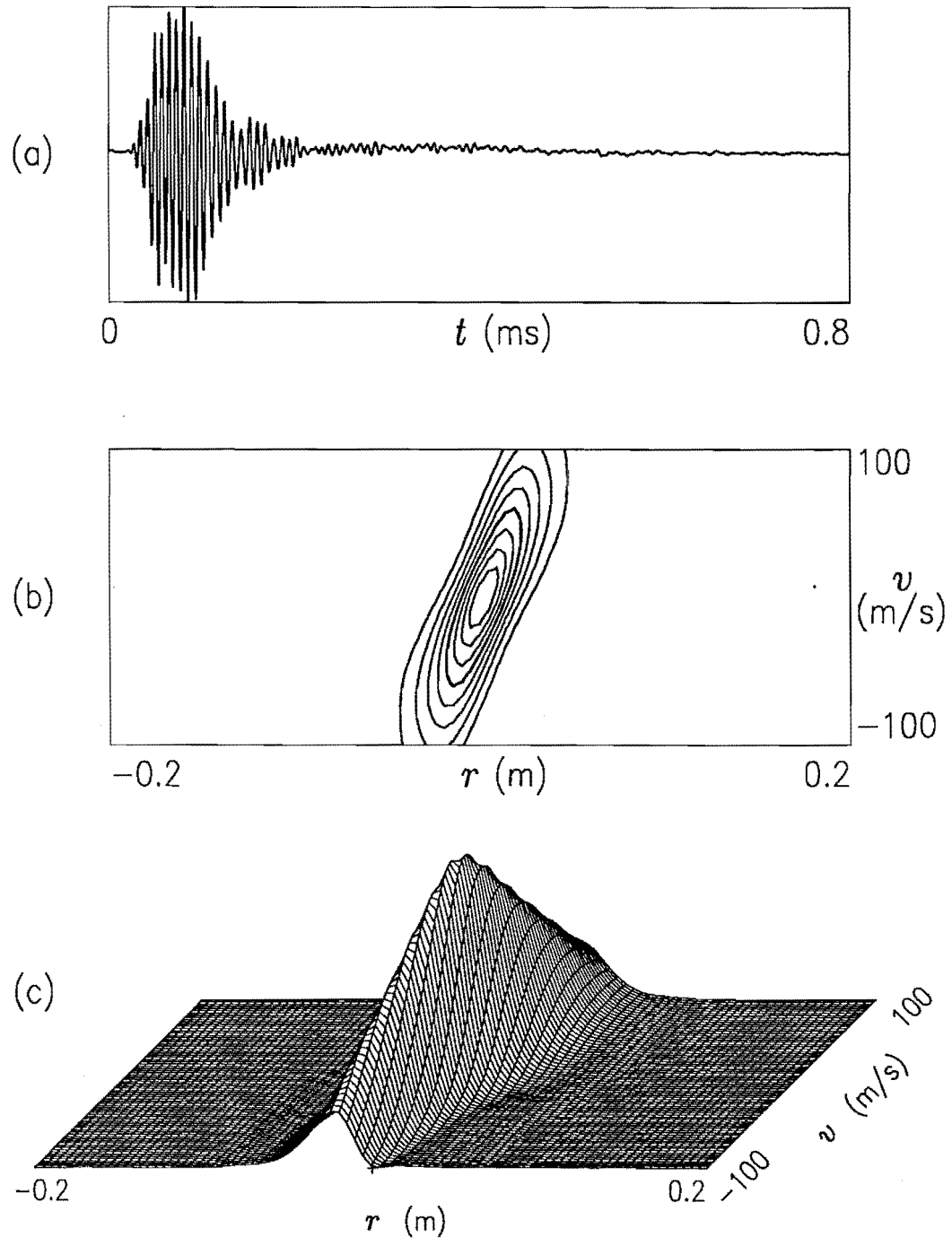


Figure 7.22. **a:** Example of a sonar click exhibiting a small frequency sweep and its ambiguity diagram: **b:** contour plot (contours spaced at $A(0,0)/10$), **c:** relief map

of velocity resolution, which is never better than 20m/s. It is extremely doubtful if any prey hunted by Hector's dolphin are capable of swimming at 20m/s (39 knots), implying that the velocity resolution of the sonar clicks is not useful for foraging. Many of the sounds indicate a slight frequency sweep, but this appears insufficient to improve velocity resolution to any useful degree.

The ambiguity diagrams of sonar sounds emitted by marine mammals have been calculated for only a few species (e.g. *Phocoena phocoena and Delphinus delphis*

Dziedzic, 1978). They mainly consist of narrow ridges aligned along the velocity axis (Dziedzic, 1978; Dziedzic *et al.*, 1977), which tend to be broader along the range axis than those of Hector's dolphins. However, Dziedzic appears to have recorded the sounds with equipment of comparatively narrow bandwidth (30kHz). More recent studies have cast doubt on the validity of the low frequency (about 2kHz) signals obtained in this way (Kamminga and Wiersma, 1981).

Hector's dolphins seem primarily concerned with resolving the ranges of targets, and their emitted sounds possess no useful doppler resolving capability. This probably stems from the high speed of sound in water, which renders the resulting doppler shifts relatively small for the range of target velocities that naturally occur. However, this velocity resolution applies only to a single click, and in no way inhibits the dolphin from inferring velocities of targets from changes in their ranges over several clicks.

Note that the ambiguity results do not infer anything about the *sensitivity* of the dolphin's sonar to target size. The range resolution of 2cm merely asserts that objects can be separately resolved if they are separated in range by more that 2cm. This does not conflict with the ability of dolphins to detect monofilament nylon down to 0.2mm in diameter (§7.1.5). The ability to detect small targets rests on the reflectivity of the target, the amount of background noise, the emitted signal strength and receiver sensitivity (§7.1.2).

With regard then to the question raised in §7.1.1 about whether Hector's dolphins can perceive gill-nets, the results presented here only provide an indication that the dolphins can resolve detail down to 2cm. Since gill-nets comprise structures that are much larger than this, it seems reasonable to assume that Hector's dolphins can perceive features of the net by means of their sonar. The previously demonstrated ability of (at least some species of) dolphins to detect monofilament nylon means that a net, which contains a large amount of such material, should offer sufficient signal strength to be detected. In addition, nets consist of floats and other supports that reflect sonar signals very well. All in all then, it seems almost certain that dolphins (of the kinds that have been studied) can detect gill-nets with their sonar. However, whether they perceive them as a danger is a completely separate question. It has also been suggested that a dolphin may not constantly employ its echo-location abilities, especially in familiar territory, and may therefore fail to detect a net even if it has the inherent ability to do so (Dawson, 19XXb; Evans *et al.*, 1988) Further discussion of the issues surrounding this question is provided by Dawson (19XXa,19XXb).

Chapter 8

Conclusions and suggestions for further research

This thesis provides a review of established techniques, and presents several new techniques, for processing sounds in order to extract some kind of useful information from them. Some concluding remarks about the new work reported in this thesis are made in §8.1, while some suggestions for how this research could be fruitfully extended appear in §8.2.

8.1 Conclusions

The preceding chapters of this thesis have discussed many different signal analysis techniques, and their application to several types of signals. Chapters 4 and 5 describe techniques that separate (speech) signals into two distinct components, based on the long-term characteristics of the signals. Chapter 6 is primarily concerned with the mechanics of implementing a signal analysis system. It also shows how spectrographic analysis allows the temporal structure of the (cough) sounds to be revealed. The analysis techniques detailed in Chapter 7 describe the structure of dolphin "clicks" in terms of simple features of their time domain and spectral envelope shapes. Statistical analyses reveal patterns in the ensemble of the sounds which are analysed.

In addition to the specific signal analysis techniques described in previous chapters, there are several "meta-themes" running through this thesis which deserve mention. The first of these is simply the observation that each analysis technique provides one with a different type of information about a signal. In order to effectively analyse a particular signal then, it is necessary to have knowledge of what is important about that signal, and of what analysis techniques are best able to characterise that particular information.

Another important point that emerges the work reported here is that useful and novel results can often be obtained through a multi-disciplinary approach to research. Both the SAA and CLEAN techniques were originally developed as methods of astronomical image reconstruction. However, as the results in Chapters 4 and 5 demonstrate, they can be successfully applied to the quite different field of speech processing. The work reported in Chapters 6 and 7 demonstrates the value of collaborating with researchers in other fields. One of the concerns of engineering is to develop practical applications of technological science, so collaboration with workers in other fields is indeed necessary in order to ensure that the most beneficial and appropriate results are obtained.

8.1.1 Speech analysis by SAA and CLEAN

The SAA technique described in Chapter 4 is a straightforward and robust approach to estimating the long-term invariant component of an utterance. The results presented in §4.3 show that the SAA signal, obtained from the voiced sections of an utterance, approximately represents the average glottal excitation signal, but with a significant contribution from the average vocal tract response. SAA signals obtained from different people differ more than those obtained from a single person on different occasions. The shape of an individual's SAA signal depends somewhat on the content of the utterance being spoken, but seems to be largely determined by the person's manner of speaking. This suggests that the SAA signal can be associated with the *vocal setting* of Laver (1980) which describes the long-term characteristics of a person's voice (see §2.1.4.1). The computational simplicity of the SAA technique means that it can be easily implemented on low-cost digital signal processing hardware (Watson *et al.*, 1988).

Chapter 5 describes the CLEAN method of subtractive deconvolution. This is applied to the problem of removing the long-term characteristics of an utterance (as represented in the SAA signal) so as to leave the variant components. In terms of the model of Laver mentioned above, the variant components can be thought of as representing the dynamic aspects of a speech signal, which correspond to the distinguishing features of different phonemes in a speech utterance (§2.1.4). CLEAN is a straightforward algorithm which iteratively identifies the positions and amplitudes of discrete pulses that represent the convolutional part of the variant component.

Because of the sparse nature of the "CLEAN" signal, the SAA and CLEAN techniques of separating a speech signal into invariant and variant components can be employed as a low data rate speech encoding scheme. The results of such a scheme, presented in §5.3, suggest that this method performs about as well as the MP-LPC method of low data rate speech encoding, especially at the medium data rates of about 16kbit/s.

In conclusion, the main advantage of the SAA/CLEAN speech analysis technique is that it provides a method of separating a speech signal into a set of components that are very different than those obtained by other analysis methods. The fact that synthetic speech can be reconstructed from these components confirms that they do represent the important features of the speech utterance. The main emphasis of the work reported in Chapters 4 and 5 is to develop the application of the SAA and CLEAN techniques to speech signals. Results are presented describing the operation and convergence of the algorithms under a wide variety of conditions.

8.1.2 Asthmatic cough sound analysis

Chapter 6 is concerned with the development of a micro-computer-based system for analysing cough sounds. This system is designed to facilitate a clinical study into the differing characteristics of cough sounds between children with and without asthma.

One of the difficulties of applying signal processing techniques to the study of clinical sounds is the large amount of data that must be managed. The COFF system, described in Chapter 6, facilitates research into cough sounds because it allows cough data from a large number of patients to be easily accessed for display and analysis.

The system employs a graphics-oriented human interface and is based on a tree-structured menu paradigm. It enables the time domain and frequency domain characteristics of the cough sounds to be interactively examined. Although the system is primarily designed to facilitate the analysis of cough sounds, it has, as far as that has been possible, been implemented in a general way so that it can easily be applied to the spectral analysis of other signals.

The system has a modular structure, which means that it is straightforward to modify or extend it. For instance, it can easily be modified to make use of other A/D boards, or its signal processing capabilities can be augmented by adding new analysis modules. §8.2.3 describes some of the ways in which it is planned to expand the system in order to automatically extract descriptive features from the cough spectrograms.

The preliminary results obtained so far (see §6.2.1) indicate that there are differences between "asthmatic" and "non-asthmatic" cough sounds. These differences are revealed by spectrographic analyses of the sounds.

8.1.3 Analysis of dolphin vocalisations

Chapter 7 describes the analysis methods invoked to obtain a quantitative description of the vocal repertoire of Hector's dolphin. The methods of automatic feature extraction facilitated the analysis of 401 0.25 second long records of Hector's dolphin click trains, containing 7661 clicks. Using basic operations in a signal processing language, 48 variables were estimated from each record. Without the aid of the computer-based analysis techniques, especially the ability to perform them automatically, this analysis would have been prohibitively labour intensive.

The feature variables were analysed by means of the principle component analysis transformation technique in order to identify the general characteristics of the clicks. The analyses performed separately on the time domain and frequency domain feature variables identified a small number of principle components that explained a large proportion of the total variance in the dataset. These components represented the number of individual peaks in the time domain or spectral envelopes respectively, together with their separations and widths. This interpretation of what features of the sounds are important is intuitively satisfying, because the occurrence and positioning of multiple peaks has relevance in terms of current models of cetacean perception. These suggest that sounds are perceived largely according to the "time-separation-pitch", which manifests itself as multiple peaks in the time domain and spectral envelopes (see §7.1.4.2; Altes, 1988; Au and Moore, 1988).

The vocal repertoire of Hector's dolphins is very simple, consisting almost entirely of short, narrowband, high frequency clicks. This is what one would expect if the clicks are used for simple echo-location purposes. However, there were also a significant minority of clicks which had a more "complicated" structure. Three possibilities suggest themselves as explanations for these signals. Firstly, they could be artefacts caused by noise, or multipath distortion. However, the high signal to noise ratio and undistorted waveforms of many examples suggest that this is not always the case. Secondly, they could be a different type of sonar signal, for use in a different type of target situation. Thirdly, they could be signals for use in communication. The relative scarcity of the "complicated" clicks in the entire sample may not indicate low importance, since one would expect most of the clicks recorded to be for sonar. High frequency cetacean sonar signals are highly directional (cf. Au *et al.*, 1986), so the strongest signals are recorded when the phonating dolphin is orientated towards the hydrophone. This situation is likely to have occurred most often when they were using sonar to examine it.

The difficulties involved in quantitative analysis and comparison of animal sounds has severely handicapped studies of acoustic behaviour. The development here of objective, automatic, feature estimation techniques has, in combination with multivariate statistical methods, facilitated a much more detailed analysis of Hector's dolphin sounds than would have been possible via previous manual methods. The development of such techniques is an important advance towards an understanding of animal sounds and their significance.

The final section of Chapter 7 is concerned with the echo-location capability

of Hector's dolphins (§7.3). Hector's dolphins, utilising the types of sounds examined in §7.3, should be capable of resolving targets down to 2cm apart. This accuracy is achieved at the expense of velocity resolution, which is never better than 20m/s. It is extremely doubtful if any prey hunted by Hector's dolphin are capable of swimming at 20m/s (39 knots), implying that the velocity resolution of the sonar clicks is not useful for foraging. Note, however, that this velocity resolution applies only to a single click, and in no way inhibits the dolphin from inferring velocities of targets from changes in their ranges over several clicks. Some of the sounds that were examined indicate a slight frequency sweep, but this appears insufficient to improve velocity resolution to any useful degree.

8.2 Suggestions for further research

This section provides details on some of the many ideas for extending the research reported in this thesis that occurred to me as I journeyed through the (rather long!) process of writing it. In addition, I present preliminary results for some practical extensions of the analysis techniques described in previous chapters. These results may encourage further research along the lines suggested.

8.2.1 Speech analysis by SAA

Several practical applications where SAA processing of speech could become useful are discussed in this section. In §8.2.1.1 I briefly describe the use of the SAA signal as a descriptor in a speaker recognition scheme. §8.2.1.2 discusses methods by which a speech signal can be "pre-processed" by the SAA signal in order to improve the subsequent estimation of descriptive parameters from the speech. These techniques attempt to remove the SAA component from the speech signal in order to leave the "variant" component only for the later processing. Note that this pre-processing is really an alternative to the method of CLEAN that is described in detail in Chapter 5. Finally, §8.2.1.3 briefly discusses the usefulness of the SAA signal as an aid in diagnosing laryngeal disorders.

8.2.1.1 Speaker recognition

As intimated by the results presented in §4.2.4.2 and §4.3, SAA can be employed to produce an estimate of a speaker's long term voice characteristics. One application where this ability is useful is in a speaker recognition scheme. As described in §3.6.2, the speaker recognition problem arises because the *intra-speaker* variation in speech characteristics is often as great as the *inter-speaker* variation. The usefulness of a particular speech descriptor for speaker recognition purposes depends on the ratios of the intra- and inter-speaker variations of that descriptor. However, even when these ratios are low, a descriptor may still be useful for speaker recognition if it characterises a different facet of voices than other descriptors. It can then be combined with those other descriptors to provide improved recognition performance.

The SAA signal characterises the long term characteristics of a speaker's voice, although, as indicated in §4.3.1.2, it does so in a different manner than does the LTAS (which has also been employed for speaker recognition, Boves, 1984, §5.3.5). Because the SAA signal does not involve a great deal of computational processing, it should be possible to combine it with other descriptors of a person's speaking style without adding too much computational complexity to the entire speaker recognition system.

As commented on in §4.2.4.2, the intra-speaker differences are greater for utterances that are composed of different phrases than those that are repetitions of the

same phrase. However, these differences are still less than the inter-speaker differences. This means that the SAA signal may be useful as a text-independent descriptor of a speaker's voice. However, SAA signals are markedly changed when a person speaks in a "different" manner (§4.2.4.2), which suggests that SAA is not useful for recognising speakers who are disguising their voices.

8.2.1.2 Improved estimation of speech parameters

In LPC analysis, the speech signal is assumed to be formed by the convolution of a train of impulses and an all-pole filter. Actual speech signals, however, contain spectral zeros, contributed by the non-impulsive glottal pulse shape and anti-resonances in the vocal tract filter. As Gray and Markel (1974) demonstrate, the numerical stability of the LPC analysis procedure is directly related to the spectral flatness of the speech signal. In order to reduce the ill effects caused by the wide dynamic range of typical speech spectra, it is usual practice to pre-emphasise the speech signal with a first-order differentiation before performing LPC analysis. This approximately counteracts the effect of non-impulsive glottal pulse shape, as is explained in §3.2.1. However, glottal pulse shapes vary between different people (cf. §3.4, §4.2.4.2), so it seems that pre-processing a speech signal with a filter derived from the SAA signal, which is characteristic of that person's glottal excitation, would be even better than merely differentiating it (Gray and Markel (1974) suggest that an adaptive pre-emphasis gives better results than a fixed one). In this section I outline the techniques involved in pre-processing speech signals with the SAA inverse filter in order to improve the LPC coefficients obtained from the speech.

One method of deconvolving the SAA signal from the speech signal is to employ Wiener filtering (§1.3.2). Figs.8.1 and 8.2 shows the spectral and time domain representations of the Wiener filter constructed from the SAA signal of the utterance AM-RAIN1, with various values for the Wiener constant ϕ. A first order differentiator is also depicted. The results of convolving the filters shown in Figs.8.2*b*,*c* and *d* onto the speech signal shown in Fig.8.3*a* appear in Figs.8.3*b*,*c* and *d* respectively. Fig.8.4 shows the Fourier magnitudes of the signals shown in the respective parts of Fig.8.3. As these figures illustrate, the effect of increasing the value of the Wiener constant is to decrease the amount of emphasis given to the high frequency components of the speech signal. This de-emphasis of the high frequencies is a well-known drawback of Wiener filtering (Bates *et al.*, 1982b) arising because the Wiener constant "swamps" the spectrum of the deconvolution kernel at those frequencies. Other techniques of deconvolution, notably CLEAN (see Chapter 5), can be invoked to "reconstruct" the higher frequency components.

Performing LPC analysis on the segments shown in Fig.8.3 and transforming the LPC coefficients to the frequency domain results in the spectra appearing in Fig.8.5. These results indicate that pre-processing speech utterances by Wiener filtering with their SAA signals may improve the estimation of LPC parameters. The spectra obtained from the LPC parameters when the speech is pre-processed with the SAA signal (especially the spectrum shown in Fig.8.5*d*) is generally flatter overall, with more prominent formant peaks, than those obtained from unprocessed and differentiated speech signals.

Note that I have only applied the processing described here to voiced sections of speech. Unvoiced sections of speech must be processed separately, either by splitting the speech signal into two sub-bands as described in Chapter 5 and then recombining them after Wiener filtering, or by performing a VUV analysis on the speech and Wiener filtering the voiced and unvoiced sections separately.

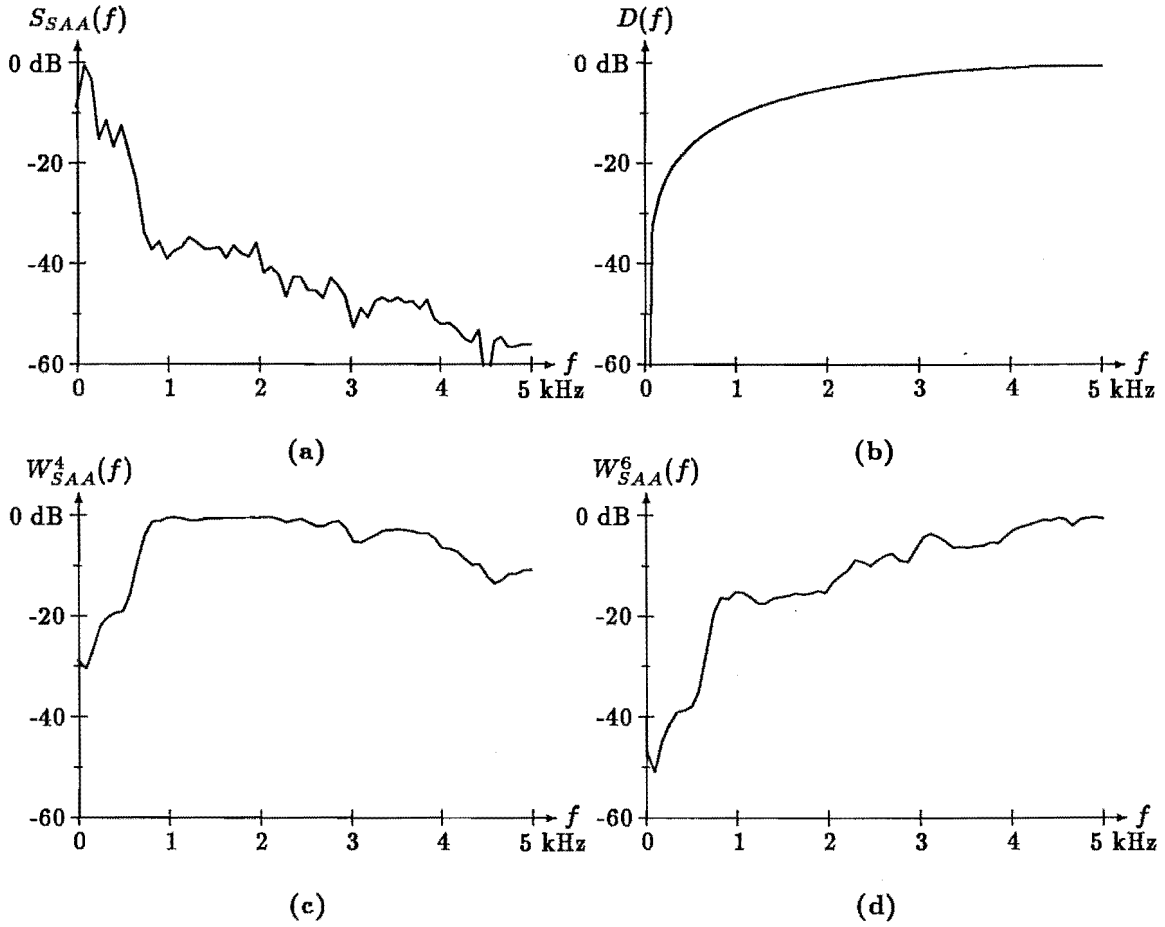


Figure 8.1. **a:** Spectrum of SAA signal of the utterance AM-RAIN1. Log spectra of **b:** a first order differentiator, **c:** and **d:** Wiener filters constructed from SAA signal shown in *a* with Wiener constants of $\phi = 10^{-4}$ and 10^{-6} respectively times the peak magnitude of $|S_{sa}(f)|^2$.

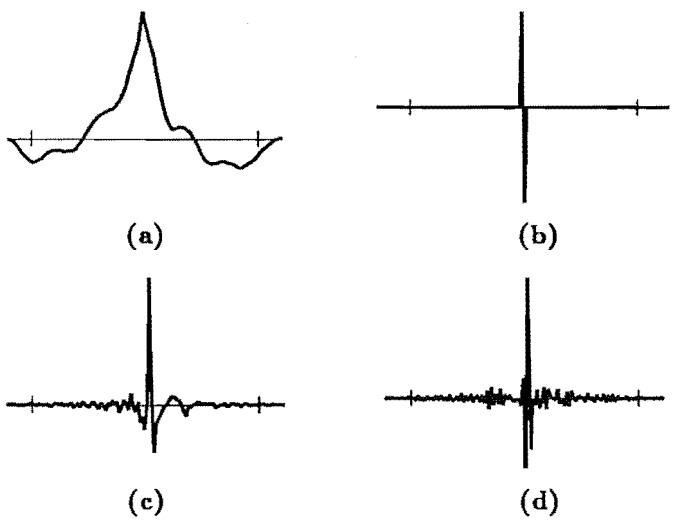


Figure 8.2. Time domain versions of the spectra shown in Fig.8.1.

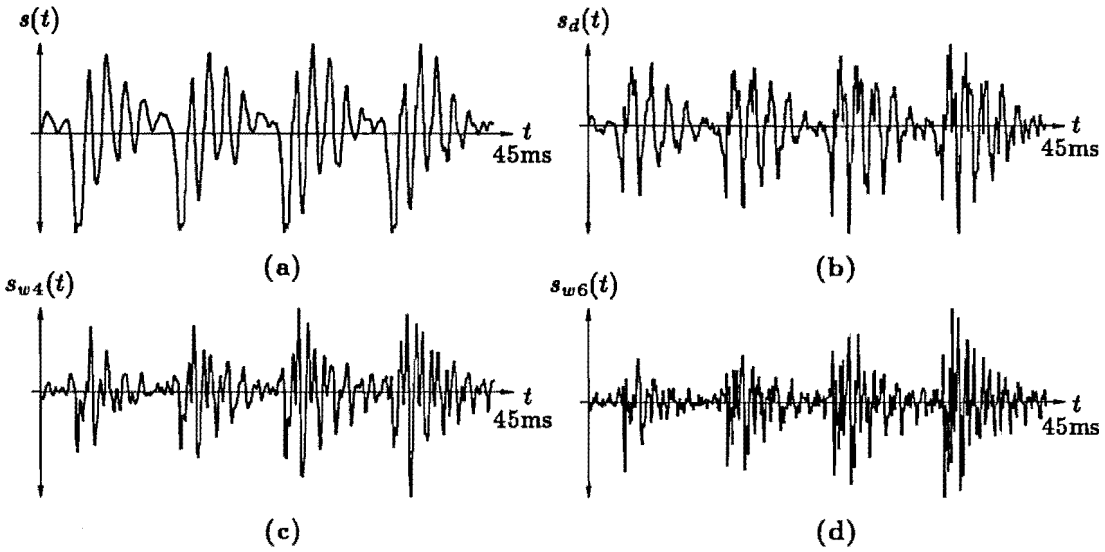


Figure 8.3. **a:** Segment of speech signal and pre-emphasised versions formed by processing with the filters shown in Figs.8.1 and 8.2. **b:** First order differentiation. **c:** and **d:** Wiener filtered versions with ϕ equal to 10^{-4} and 10^{-6} times the peak of $|S_{aa}(f)|^2$ respectively.

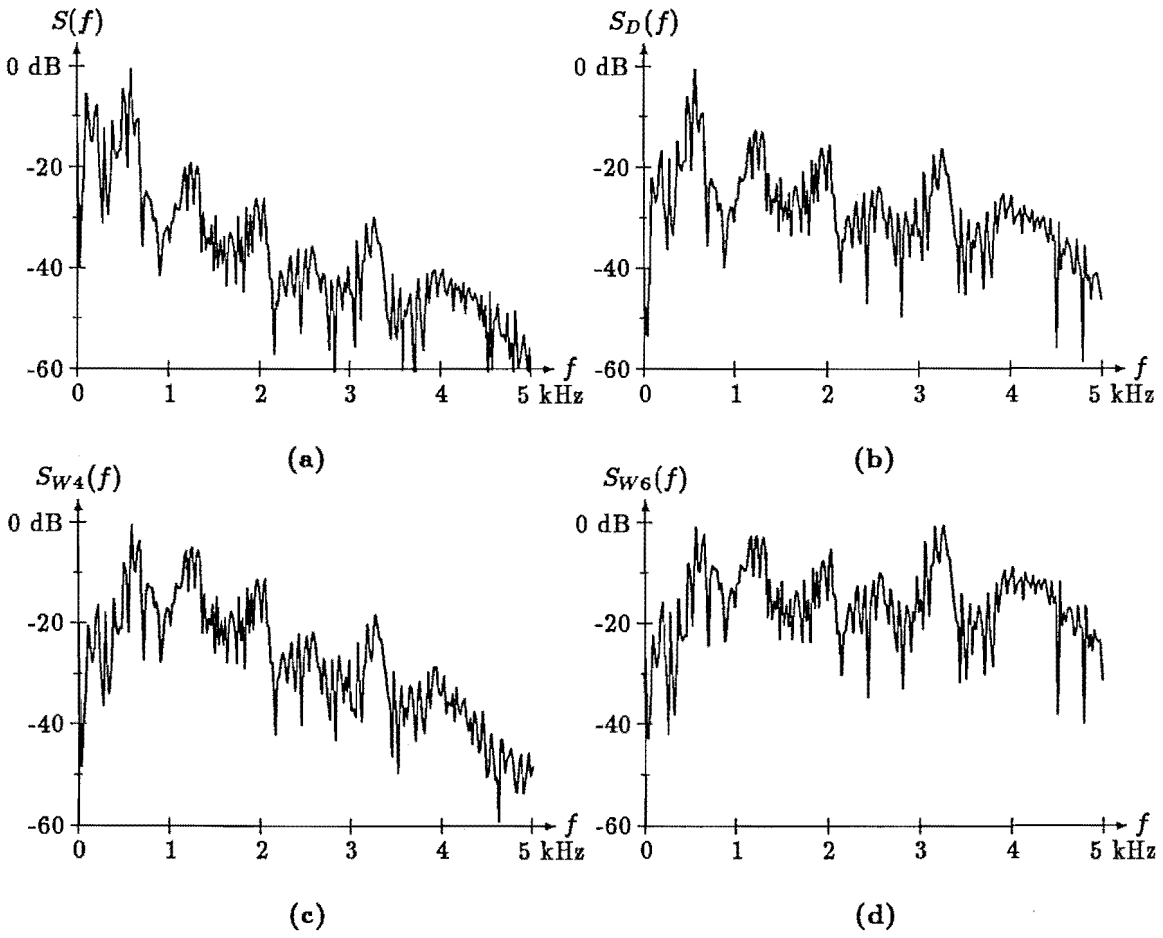


Figure 8.4. Fourier magnitudes of the signals shown in Fig.8.3. **a:** Speech signal $|S(f)|$. **b:** Pre-emphasis by first order differentiation. **c:**, **d:** SAA signal deconvolved out with Wiener constant of 10^{-4} and 10^{-6} times the peak of $|S_{aa}(f)|^2$ respectively.

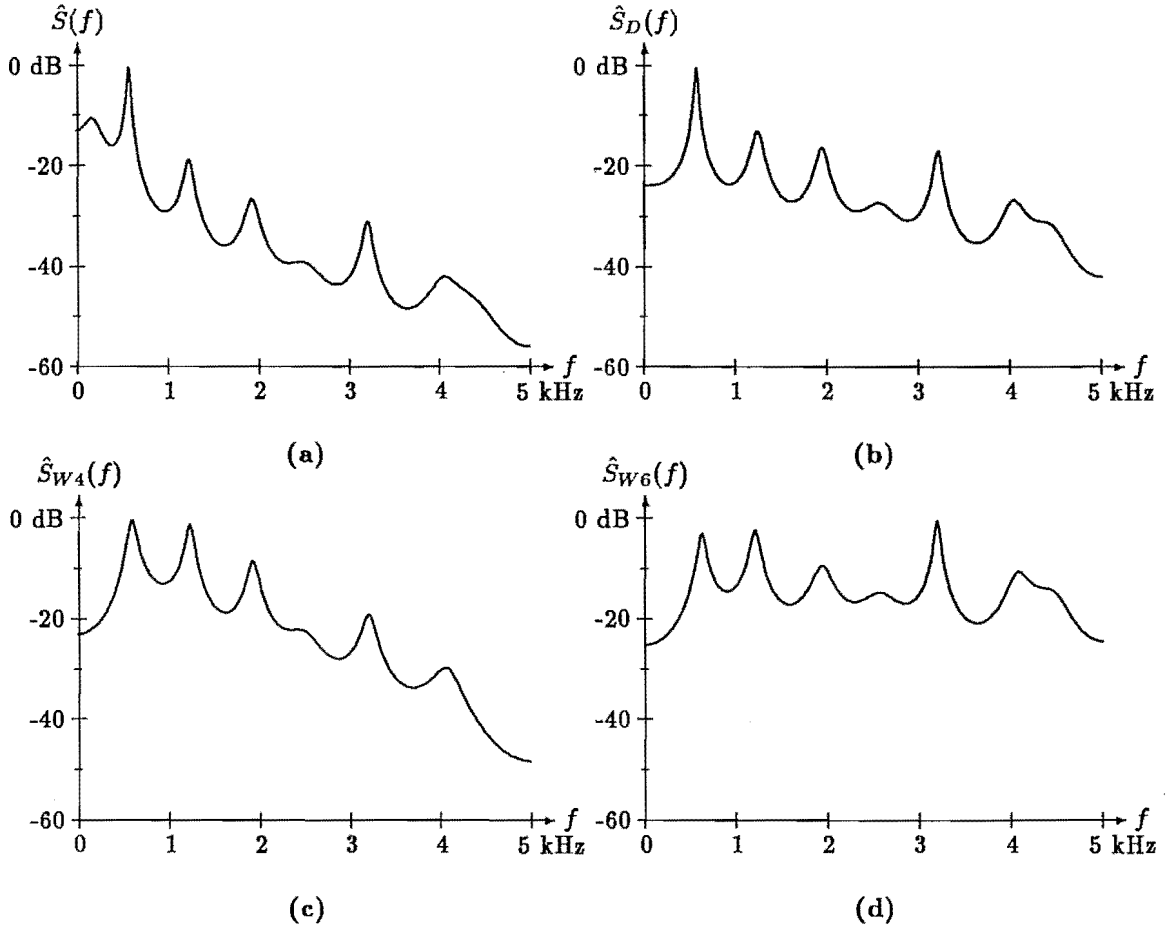


Figure 8.5. Spectra (computed via LPC analysis) of the segments of deconvolved speech shown in the respective parts of Fig.8.3. Sixteen LPC coefficients were computed from each segment.

8.2.1.3 Investigating glottal source characteristics

In order to investigate the possibility of using the SAA technique to characterise the state of a speaker's larynx, I processed utterances spoken in "tense", "relaxed", and "normal" manners. These different ways of speaking change the glottal excitation because they change the muscle tension around the larynx. Although these results (see Fig.4.11 in §4.2.4.2) do not reveal anything about specific laryngeal disorders, they indicate that the SAA signal changes markedly in response to changes in the larynx. Together with the results in §4.3 showing that the SAA signal can provide an estimate of the average glottal excitation (albeit combined with the average vocal tract response), these results suggest that the SAA signal may provide an indication of those laryngeal pathologies that change the shape of the glottal waveform (cf. §3.6.4.1). Further research is required to investigate the differences in the shapes of SAA signals to be expected from speakers with various laryngeal disorders.

In addition to the possible uses as a diagnostic aid, the ability to characterise the glottal excitation is important in endeavours to synthesise "natural-sounding" speech (cf. Gauffin and Sundberg, 1989). In order to evaluate the usefulness of SAA in this role, I recommend that a comprehensive investigation be carried out into the changes in the SAA signal that occur when a person talks in different ways. It would be useful to compare the results of such an investigation with those obtained by other methods of glottal source characterisation (cf. Gauffin and Sundberg, 1989; Gobl, 1989).

8.2.2 Speech analysis by CLEAN

There are two main areas in which the CLEAN technique described in Chapter 5 could be further developed. Firstly, the method employed to estimate the pulse positions and amplitudes could be improved. §8.2.2.1 and §8.2.2.2 describe two approaches for seeking such an improvement. The second area where useful gains could be made is in the encoding of the pulses. §8.2.2.3 makes some suggestions in this regard.

In addition to the low data rate speech encoding scheme described in Chapter 5, there are several other applications where SAA/CLEAN speech analysis could be useful. Several of these are briefly discussed in §8.2.2.4.

8.2.2.1 Refinement of pulse positions

The optimisation procedure described in §5.2.6 only optimises the amplitudes of the pulses that have been found by the CLEAN algorithm — it does not adjust their positions. In this section I describe the iterative position refinement scheme introduced by Bates *et al.* (1982c) and briefly examine its usefulness as a method of improving CLEAN processing of speech signals. This scheme individually adds the CLEAN kernel corresponding to each of the CLEAN pulses to the residual signal and then re-estimates the position and amplitude of that pulse.

The residual dirty signal $r(t)$ remaining after CLEAN has been performed on a speech signal $s(t)$ can be expressed as

$$r(t) = s(t) - \sum_{i=1}^{N_p} v_i g(t - p_i) \tag{8.1}$$

where $g(t)$ is the CLEAN kernel and the v_i and p_i are the amplitudes and positions respectively of the N_p pulses in the CLEAN signal. Now, for $j = 1 \ldots N_p$, the j^{th} pulse is removed from the CLEAN signal and the appropriately weighted kernel is added to the residual. The j^{th} *interim residual* $\xi_j(t)$ is then given by

$$\xi_j(t) = r_{j-1}(t) + v_j g(t - p_j), \tag{8.2}$$

where $r_0(t) = r(t)$. The position and amplitude of the j^{th} pulse is then re-estimated as indicated in Steps 1 and 2 of the CLEAN algorithm described in §5.2.3. The residual signal is updated by

$$r_j(t) = \xi_j(t) + \hat{v}_j g(t - \hat{p}_j), \tag{8.3}$$

where $\hat{v}_j$ and $\hat{p}_j$ are the refined pulse amplitude and position respectively.

Fig.8.6*a* shows the changes in SNR that result when the postions of the pulses in a CLEAN signal are refined by the above procedure. The SNR increases at first, but then decreases as further refinement iterations are applied (where a "refinement iteration" refers to the application of the above procedure to every individual CLEAN pulse). The curves shown in Fig.8.6*b* indicate that the number of CLEAN pulses is reduced slightly after each refinement iteration. This occurs because closely-spaced pulses may be combined by the refinement procedure.

It would be worthwhile to investigate further how much improvement in the accuracy of the CLEAN signal can be obtained by this procedure, and whether it justifies the additional processing that is required. In addition, the ability to improve the positions of the pulses may be a useful pre-processing step for the technique of estimating the vocal tract area from the CLEAN signal postulated in §5.4.1.3.

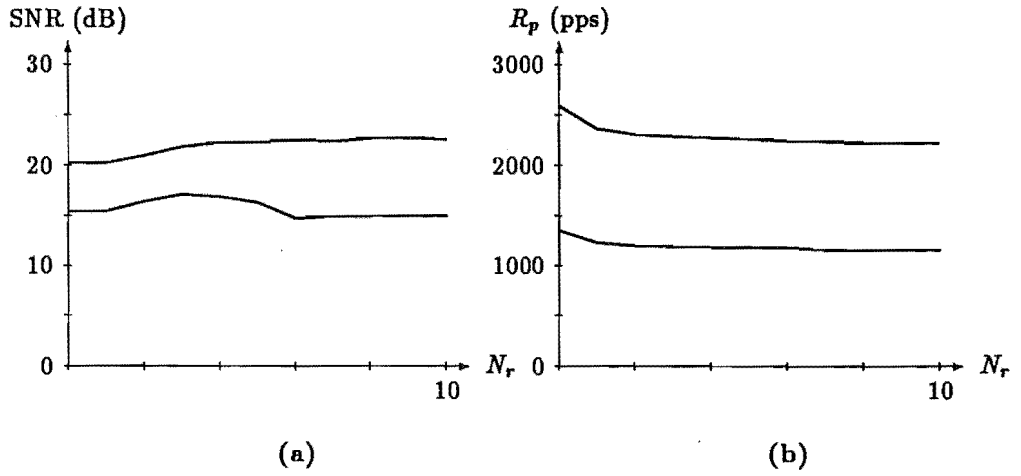


Figure 8.6. **a:** SNR, and **b:** pulse rate, versus the number of iterations of the position refinement scheme. The two lines represent two different CLEAN signals (of the speech segment shown in Fig.5.3*a*, with pulse rates of 1000 and 2000ppsrespectively).

8.2.2.2 Locating pulses by cross-correlation

In the basic CLEAN algorithm presented in §5.2.3, the j^{th} pulse is located at the instant for which $|r^{(j)}[n]|$ is a maximum. A more "optimal" location can be found by minimising the mean-square level of the residual $E^{(j+1)}$ at the $(j+1)^{\text{th}}$ iteration, where $E^{(j+1)}$ is defined by

$$E^{(j+1)} = \sum_{n=-\infty}^{\infty} \left[r^{(j)}[n] - v_j g[n - p_j]\right]^2. \tag{8.4}$$

Referring to (5.15) through (5.19) in §5.2.6, replacing $y[n]$ with $r^{(j)}[n]$, and setting $N_{\text{opt}} = 1$, the "optimal" value for v_j is given by

$$v_j = c_{p_j} / g_{p_j p_j}. \tag{8.5}$$

By substituting (8.5) into (8.4) and recalling the definitions of c_i and g_{ij} given in §5.2.6, the error $E^{(j+1)}$ can be expressed as a function of possible pulse positions i by means of

$$E^{(j+1)}[i] = \sum_{n=-\infty}^{\infty} r^{(j)}[n]^2 - c_i^2 / g_{ii}. \tag{8.6}$$

The "optimal" pulse position is defined as the value of i which minimises $E^{(j+1)}[i]$, which, since only the second term on RHS (8.6) depends on i, is given by the maximum of c_i^2 / g_{ii}.

Although steps 1 and 2 of the algorithm described in §5.2.3 can be modified to implement (8.6) and (8.5) directly, the cross-correlation c_i between $g[n]$ and $r^{(j)}[n]$ must then be evaluated at each iteration, which means that the computational requirements are greatly increased. It turns out, however, that the computational load can be reduced considerably by invoking the following recursive scheme. Initially, $c_i^{(1)}$ is set to be the cross-correlation between $g[n]$ and the speech signal $s[n]$. The j^{th} pulse is found at the maximum of $|c_i^{(j)}|$ (since g_{ii} is a constant), where, for $j > 1$,

$$c_i^{(j)} = c_i^{(j-1)} - v_{j-1} g_{p_{j-1} i}. \tag{8.7}$$

The amplitude of v_j is given by (8.5), evaluated with $c_{n_j}^{(j)}$. Since the cross-correlation is updated at each step, it does not need to be re-evaluated. This is in fact the same scheme that is invoked to locate the pulses in the multi-pulse LPC technique (see §3.5.2.4 for the details).

It is interesting to compare this recursive pulse locating scheme with the CLEAN algorithm described in §5.2.3. Inspection of (8.5) and (8.7), recalling that $g_{ij} = \mathrm{gg}[|i-j|]$, reveals that this scheme is actually equivalent to the CLEAN algorithm, but with $s[n]$ replaced by $c_i^{(1)}$ and $\mathrm{gg}[i]$ taking the place of $g[n]$ (provided that $\mathrm{gg}[0] = 1$). Hence the simplicity of the CLEAN algorithm still applies to the "optimal" pulse location scheme.

Some informal experiments with this technique of locating the pulses suggested that it did not seem to greatly improve on the performance of the basic CLEAN procedure described in §5.2.3. However, if it turns out that the re-optimisation step that is invoked in the "standard" approach is not necessary with this modified approach, a significant saving in computation would accrue.

I think that it is also worthwhile to pursue this approach because of more fundamental considerations. As discussed in detail in §5.2.2, the assumptions that underly the use of $\max|r(t)|$ as an estimate for the CLEAN pulse position and amplitude (i.e. that $g(t)$ is "peaky") are not fully satisfied by speech signals. Employing the maximum cross-correlation between $r(t)$ and $g(t)$ should provide a better estimate of the pulse amplitude and position because it takes into account the overall shape of $g(t)$. Further research along this line could also be useful in some of the other areas in which CLEAN has been applied (Note that Högbom initially considers the use of the cross-correlation criterion, but is able to resort to the more straightforward maximum magnitude criterion because the nature of the radio-astronomical images means that the two approaches are equivalent).

8.2.2.3 Modifications to the SAA/CLEAN low data rate speech encoding scheme

Chapter 5 describes a low data rate speech encoding scheme which employs the SAA and CLEAN analysis techniques. However, this speech encoding scheme was implemented in a signal processing development environment on a multi-user computer system. Furthermore, in order to simplify the development of the scheme, it was designed so that the component operations of the scheme were carried out on the entire utterance in sequential order. This type of implementation is suited to applications when the entire utterance is available throughout the processing, such as for compression of speech data files on a computer system. For real-time operation, such as is required for low data rate speech transmission, the entire sequence of operations must be carried out on short segments of the input speech signal, so as not to introduce undue delay into the signal path. In particular, the SAA signal cannot be obtained over the whole utterance being encoded, as is done in §5.3. The results presented in Chapter 4 (see §4.2.4.2) indicate that, providing a person continues to talk in a similar way, their SAA signal remains fairly consistent. The SAA signal obtained from a "training period" should therefore be suitable for use in a subsequent period. A useful line of future research would be to implement the scheme in a real-time environment, so that its performance as a practical method of low data rate speech transmission can be evaluated.

The low data rate of the SAA/CLEAN scheme is a consequence of the sparseness of the CLEAN signal. However, its coding efficiency is very dependent on the efficiency with which the CLEAN pulses are encoded into binary numbers. One suggestion to improve the efficiency of the pulse encoding is to combine the pulse sequences

from the two sub-bands and encode them jointly. Because the statistics of the joint sequence are more regular than those of the individual sequences on their own, a reduced data rate should result. Note however, that it is probably not worthwhile trying to implement an encoding scheme that is too complicated, because that would erode the computational advantage that SAA/CLEAN has over some other speech coders.

8.2.2.4 Miscellaneous applications of CLEAN speech analysis

In addition to the low data rate speech coding scheme referred to in §8.2.2.3, there are several applications where it would be useful to investigate the use of SAA/CLEAN speech analysis.

One such application is to use the CLEAN pulses as a feature in a recognition system. This may be useful because, as indicated in §5.4, the CLEAN signal represents the linguistic information in an utterance in a somewhat different fashion than the LPC coefficients. Two approaches to putting the CLEAN pulses in a form suitable for recognition are suggested here. One approach is the use of CLEAN as a pre-processing step in a standard LPC analysis, in a similar manner to the use of Wiener filtering described in §8.2.1.2. The purpose of CLEAN in this approach is therefore to remove the contribution of the SAA signal from the speech utterance, therefore improving the estimated LPC coefficients. It would be interesting to compare the results obtained by pre-processing in this way with CLEAN with those obtained via the Wiener filtering approach outlined in §8.2.1.2. Other comparisons of CLEAN and Wiener filtering have noted that each method is suited to different types of signals (see §5.1.3; Bates *et al.*, 1982b,1984).

Another approach to abstracting features from the CLEAN signal is to employ the CLEAN pulses themselves as feature parameters. However, this is not a straightforward undertaking, both because of the varying number of pulses from segment to segment, and because both the pulse positions and amplitudes are significant. In addition, the pulses are not "robust" in the sense that small changes in the signal may induce large changes in the CLEAN pulses. For instance, one large pulse may be nearly equivalent (as far as the CLEAN algorithm is concerned) to two adjacent pulses of smaller amplitude. However, in terms of feature paramaters, two smaller, adjacent, pulses cannot be easily reconciled with a single large pulse. Possibly the pulse position refining technique introduced in §8.2.2.1 would help to alleviate this difficulty.

In order to compare the CLEAN pulses from two speech sounds, the pulses must be somehow "aligned". One suggestion is to extract fixed numbers of pulses from each segment in a "pitch -synchronous" fashion. As argued in §5.4.1.3, the largest pulse in a pitch period can be thought of as the "primary excitation" pulse. It can therefore serve as a reference for the remaining pulses in the pitch period. The feature vector is then composed of the relative positions and amplitudes of these pulses, which can be thought of as representing reflections in the vocal tract. However, it is not clear how to order the pulses in the feature vector. They could be arranged in order of magnitude, in which case different vectors are distinguished largely by the positions of each successively smaller pulse. The other way to arrange them is in their temporal order. In this approach, the vectors are distinguished by the amplitude and position of each successive pulse.

Another possible use for CLEAN speech analysis could be as part of a method of estimating the vocal tract shape area. As discussed in §5.4.1.3, the CLEAN signal can be associated with the vocal tract impulse response if the CLEAN kernel is assumed to represent the glottal excitation. Since the SAA signal does not really fulfil the latter assumption, as is indicated by the results in §4.3, it would be interesting to perform some experiments with synthetic speech (where the excitation shape is known) to find

out how closely the CLEAN signal matches the impulse response of a discrete tube model of the vocal tract. In this regard, it may also be worthwhile to see if the position refinement (§8.2.2.1) and cross-correlation pulse identification (§8.2.2.2 schemes provide better estimates of the pulse positions in the impulse response. If the method is applied to natural speech, it may be necessary to employ estimates of the glottal excitation other than the SAA signal in order to obtain a better estimate of the vocal tract impulse response.

8.2.3 Asthmatic cough sound analysis

The work reported in Chapter 6 primarily concerns the development of a clinical cough sound collection, analysis, and data management system. This is part of a wider research effort to identify changes in the characteristics of cough sounds that occur as a result of asthma. In this section I present some of the ways in which the COFF system is being expanded so as to extract quantitative features from the cough sounds.

The goal of the research project for which the COFF system was implemented is to characterise the differences between the coughs of asthmatic and non-asthmatic children. The COFF system as described in Chapter 6 facilitates the collection of cough sounds, and allows their temporal and spectral structure to be interactively examined.

In order to pursue the development of quantitative analysis techniques which characterise the cough sounds, a framework is required in which candidate analysis techniques can be implemented and their usefulness evaluated. What I propose is a "shell" program into which specific analysis procedures can easily be inserted. The shell program would implement all the details of getting the cough data from the appropriate files and of collating the results of the analysis. Initially it would be worthwhile to make use of a stand-alone statistical analysis package to investigate the results of the analysis. The shell program would therefore need to create data files in a format that can be imported into the statistical package. When a set of useful feature variables has finally been chosen, the COFF system could be extended to implement the appropriate analysis procedures.

As far as the actual analysis procedures are concerned, the preliminary results of §6.2.1 and those of other researchers discussed in §6.2.3 suggest that it will first be necessary to separate each cough into several segments representing the different "phases" that can be observed in a typical cough. This segmentation is most simply performed by applying thresholds to the energy envelope of the sound to identify the beginning and end points of the segments. If this approach proves unreliable, one of the many adaptive segmentation techniques (cf. Bodenstein and Praetorius, 1977; Andre-Obrecht, 1988) could be applied. Adaptive segmentation may be necessary because of the confusion that could otherwise arise between a loud "second phase" and the final burst itself (see §6.2.3).

Features that could be extracted from the coughs range from simple descriptors such as the mean frequency, energy, and duration of each of the segments referred to above, to more sophisticated shape descriptors such as LPC or cepstral coefficients. It could also be useful to identify and track spectral peaks in order to recognise the occurrence of "wheezes" in the cough sound.

8.2.4 Analysis of dolphin vocalisations

There are several useful directions in which to extend the research described in Chapter 7 of characterising the vocalisations of hector's dolphins. Firstly, it would be interesting to analyse longer sequences of clicks. Practical considerations limited the duration of the click sequences that were analysed to 0.25 s. However, when listening

to recordings of the clicks, there sometimes appear to be patterns in the click sequence over much longer time spans. These patterns are revealed to the ear as varying tones caused by changes in the click rate. It would be worthwhile to digitise a few of these longer click sequences and analyse the long-term structure of the click sequence.

Another useful extension of this research would be to apply similar techniques to the analysis of other animal vocalisations. As mentioned in §7.1.5, digital analysis techniques are only beginning to be applied to studies of animal sounds. In order for such methods to become widely available to zoologists, it is necessary that a range of analysis techniques be developed, suitable for characterising the wide variety of animal sounds that are encountered. The analysis techniques described in Chapter 7 are suitable for "impulsive" types of sounds which can be described by the shapes of their time domain and spectral envelopes. For longer duration signals (such as dolphin whistles), such techniques are less suitable because of the time-varying nature of the signal. It is necessary to segment the signal and describe it in terms of the temporal variation in its short-term characteristics. Suitable analysis techniques include the use of the time-varying spectrum (§3.3.1), adaptive segmentation (cf. Bodenstein and Praetorius, 1977), and syntactic descriptions of the temporal structure (cf. Fu, 1974). The use of temporal normalisation schemes such as those employed in speech recognition (§3.6.1) may also be necessary in order to take into account small variations in the durations of different signals.

References

AKASAKA, K, KONNO, K, ONO, Y, MUE, S, ABE, C, KUMAGAI, M and ISE, T (1975), 'Acoustical studies on respiratory sounds in asthmatic patients', *Tohoku J. Exp. Med.*, Vol. 117, pp. 323–333.

ALLEN, J (1975), 'Computer architecture for signal processing', *Proc. IEEE*, Vol. 63, No. 4, pp. 624–633.

ALLEN, JB (1985), 'Cochlear modeling', *IEEE ASSP Magazine*, Vol. 2, No. 1, pp. 3–29.

ALMEIDA, LB and TRIBOLET, JM (1984), 'Harmonic coding: An introduction', In DEWILDE, P and MAY, CA (Eds.), *Int. Conf. on Commun.*, IEEE, Elsevier, pp. 1169–1173.

ALTES, RA (1988), 'Some theoretical concepts for echolocation', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 725–752.

ANANTHAPADMANABHA, TV and FANT, G (1982), 'Calculation of the true glottal flow and its components', *Speech Communication*, Vol. 1, pp. 167–184.

ANANTHAPADMANABHA, TV and YEGNANARAYANA, B (1979), 'Epoch extraction from linear prediction residual for identification of closed glottis interval', *IEEE Trans. ASSP*, Vol. 27, No. 4, pp. 309–319.

ANASTAPLO, S and KARNELL, MP (1988), 'Synchronized videostroboscopic and electroglottographic examination of glottal opening', *J. Acoust. Soc. Am.*, Vol. 83, No. 5, pp. 1883–1890.

ANDRE-OBRECHT, R (1988), 'A new statistical approach for the automatic segmentation of continuous speech signals', *IEEE Trans. ASSP*, Vol. 36, No. 1, pp. 29–40.

ANDREWS, HL (1984), 'Speech processing', *Computer*, Vol. 17, No. 10, pp. 315–324.

ANONOMOUS (1988), 'Cough and wheeze in asthma: are they interdependent', *Lancet*, Vol. 1, No. 8583, pp. 447–448, (Editorial).

ARASEKI, T, OZAWA, K, ONO, S, YASUNAGA, S, WAKE, Y and TANAKA, S (1986), 'A high quality multi-pulse LPC coder for speech transmission below 16kbps', In *Int. Conf. on Digital Satellite Communications-7*, Munich, pp. 785–790.

ARCHER, LNJ and SIMPSON, H (1985), 'Night cough counts and diary scores in asthma', *Arch. Dis. Child.*, Vol. 60, No. 5, pp. 473–474.

ARONSON, AE (1985), *Clinical Voice Disorders*, Thieme, New York, 2nd ed.

ARRILLAGA, J, BRADLEY, DA and BODGER, PS (1985), *Power System Harmonics*, John Wiley, Chichester.

ATAL, BS (1972), 'Automatic speaker recognition based on pitch contours', *J. Acoust. Soc. Am.*, Vol. 52, No. 6 (Pt. 2), pp. 1687–1697.

ATAL, BS (1974), 'Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification', *J. Acoust. Soc. Am.*, Vol. 55, No. 6, pp. 1304–1312.

ATAL, BS (1985), 'Linear predictive coding of speech', In FALLSIDE, F and WOODS, WA (Eds.), *Computer Speech Processing*, Prentice Hall, New Jersey.

ATAL, BS and HANAUER, SL (1971), 'Speech analysis and synthesis by linear prediction', *J. Acoust. Soc. Am.*, Vol. 50, No. 2 (part 2), pp. 637–655.

ATAL, BS and RABINER, LR (1986), 'Speech research directions', *AT & T Tech. J.*, Vol. 65, No. 5, pp. 75–88.

ATAL, BS and REMDE, JR (1982), 'A new model of LPC excitation for producing natural-sounding speech at low bit rates', In *Int. Conf. on ASSP*, IEEE, pp. 614–617.

ATAL, BS and SCHROEDER, MR (1968), 'Predictive coding of speech signals', In *The 6th Int. Cong. on Acoustics*, Tokyo, Japan, pp. C13–C16.

ATAL, BS and SCHROEDER, MR (1970), 'Adaptive predictive coding of speech signals', *Bell Syst. Tech. J.*, Vol. 49, No. 8, pp. 1973–1986.

ATAL, BS and SCHROEDER, MR (1979), 'Predictive coding of speech signals and subjective error critera', *IEEE Trans. ASSP*, Vol. 27, No. 3, pp. 247–254.

ATAL, BS and SCHROEDER, MR (1984), 'Stochastic coding of speech signals at very low bit rates', In DEWILDE, P and MAY, CA (Eds.), *Int. Conf. on Commun.*, IEEE, Elsevier, pp. 1610–1613.

AU, WWL (1979), 'Echolocation signals of the Atlantic Bottlenose dolphin (*Tursiops truncatus*) in open waters', In BUSNEL, R and FISH, JF (Eds.), *Animal Sonar Systems*, Plenum Press, New York, pp. 251–282.

AU, WWL (1988a), 'Detection and recognition models of dolphin sonar systems', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 753–768.

AU, WWL (1988b), 'Sonar target detection and recognition by odontocetes', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 451–465.

AU, WWL and MOORE, PWB (1988), 'The perception of complex echoes by an echolocating dolphin', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 295–299.

AU, WWL and PAWLOSKI, JL (1989), 'Detection of noise with rippled spectra by the Atlantic bottlenose dolphin', *J. Acoust. Soc. Am.*, Vol. 86, No. 2, pp. 591–596.

AU, WWL, MOORE, PWB and PAWLOSKI, D (1986), 'Echolocation transmitting beam of the Atlantic bottlenose dolphin', *J. Acoust. Soc. Am.*, Vol. 80, No. 2, pp. 668–691.

AVITAL, A, BAR-YISHAY, E, SPRINGER, C and GODFREY, S (1988), 'Bronchial provocation tests in young children using tracheal auscultation', *J. Pediatr.*, Vol. 112, No. 4, pp. 591–594.

AYERS, GR and DAINTY, JC (1988), 'An iterative blind deconvolution algorithm and its applications', *Opt. Lett.*, Vol. 13, No. 7, pp. 547–549.

BAKER, AN (1978), 'The status of Hector's dolphin, *Cephalorhynchus hectori* (Van Beneded) in New Zealand waters', *Rep. Int. Whal. Commn.*, Vol. 28, pp. 331–334.

BANCI, G, MONINI, S, FALASCHI, A and DE SARIO, N (1986), 'Vocal fold disorder evaluation by digital speech analysis', *J. Phon.*, Vol. 14, pp. 495–499.

BATES, RHT (1976), 'A stochastic image restoration procedure', *Opt. Commun.*, Vol. 19, No. 2, pp. 240–244.

BATES, JHT (1981), *Applications of modelling and image processing in medicine*, PhD thesis, University of Otago, Dunedin, N.Z.

BATES, RHT (1982), 'Astronomical speckle imaging', *Physics Reports*, Vol. 90, No. 4, pp. 203–297.

BATES, RHT (1987), 'Some image processing: Highlights in retrospect', *Search*, Vol. 18, No. 5, pp. 237–240.

BATES, RHT and CADY, FM (1980), 'Towards true imaging by wideband speckle interferometry', *Optics Communication*, Vol. 32, pp. 365–369.

BATES, RHT and DAVEY, BLK (1987), 'Towards making shift-and-add a versatile imaging technique', In IDELL, PS (Ed.), *Proc. SPIE: Vol. 828: Digital Image Recovery and Synthesis*, pp. 87–94.

BATES, RHT and MCDONNELL, MJ (1986), *Image Restoration and Reconstruction*, Oxford University Press, Oxford.

BATES, RHT and ROBINSON, BS (1981), 'Ultrasonic transmission speckle imaging', *Ultrasonic Imaging*, Vol. 3, No. 4, pp. 378–394.

BATES, RHT and ROBINSON, BS (1982), 'A stochastical imaging procedure', In ASH, EA and HILL, CR (Eds.), *Acoustical Imaging*, Plenum, pp. 185–191.

BATES, JHT, FRIGHT, WR, MILLANE, RP, SEAGAR, AD, BATES, GTH, NORTON, WA, MCKINNON, AE and BATES, RHT (1982a), 'Subtractive image restoration. III. Some practical applications', *Optik*, Vol. 62, No. 4, pp. 333–346.

BATES, JHT, MCKINNON, AE and BATES, RHT (1982b), 'Subtractive image restoration. II: Comparison with multiplicative deconvolution', *Optik*, Vol. 62, No. 1, pp. 1–14.

BATES, JHT, MCKINNON, AE and BATES, RHT (1982c), 'Subtractive image reconstruction. I: Basic theory', *Optik*, Vol. 61, No. 4, pp. 349–364.

BATES, JHT, FRIGHT, WR and BATES, RHT (1984), 'Wiener filtering and cleaning in a general image processing context', *M. Not. R. Astron. Soc.*, Vol. 211, pp. 1–14.

BATES, RHT, BRIESEMAN, NP, CLARK, TM, ELDER, AG, FRIGHT, WR, GARDEN, KL, KENNEDY, WK, SQUIRES, PL, THORPE, CW, TURNER, SG and JELINEK, HJ (1987), 'Interactive speech-defect diagnostic/therapeutic/prosthetic aid', In LETELLIER, JP (Ed.), *Proc. SPIE: Vol. 827: Real Time Signal Processing X*, pp. 131–139.

BAUGHMAN, RP and LOUDON, RG (1985), 'Lung sound analysis for continuous evaluation of airflow obstruction in asthma', *Chest*, Vol. 88, pp. 364–368.

BECK, R and GAVRIELY, N (1990), 'The reproducibility of forced expiratory wheezes', *Am. Rev. Respir. Dis.*, Vol. 141, pp. 1418–1422.

BEHRENS, SJ and BLUMSTEIN, SE (1988), 'Acoustic characteristics of English voiceless fricatives: a descriptive analysis', *J. Phon.*, Vol. 16, pp. 295–298.

VAN DEN BERG, JW (1968), 'Mechanism of the larynx and the laryngeal vibration', In MALMBERG, B (Ed.), *Manual of Phonetics*, North-Holland, Amsterdam, Chap. 9, pp. 278–308.

VAN DEN BERG, JW, ZANTEMA, JT and DOORNENBAL, JR., P (1957), 'On the air resistance and the Bernoulli effect of the human larynx', *J. Acoust. Soc. Am.*, Vol. 29, No. 5, pp. 626–631.

BERNSTEIN, LE, GOLDSTEIN, MH and MAHSHIE, JJ (1988), 'Speech training aids for hearing-impaired individuals: I Overview and aims', *J. Rehab. Res. Dev.*, Vol. 25, No. 4, pp. 53–62.

BICKLEY, CA and STEVENS, KN (1986), 'Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies', *J. Phon.*, Vol. 14, pp. 373–382.

BINGHAM, C, GODFREY, MD and TUKEY, JW (1967), 'Modern techniques of power spectrum estimation', *IEEE Trans. Audio Electroacoust.*, Vol. 15, No. 2, pp. 56–66.

BIRNBAUM, M, COHEN, LA and WELSH, FX (1986), 'A voice password system for access security', *AT & T Tech. J.*, Vol. 65, No. 5, pp. 68–74.

BLACKMAN, RB and TUKEY, TW (1958), *The Measurement of Power Spectra*, Dover, New York. Also publ. in *Bell Syst. Tech. J.*, Vol. 37, January and March 1958).

BLADON, A (1983), 'Two-formant models of vowel perception: Shortcomings and enhancements', *Speech Communication*, Vol. 2, No. 4, pp. 305–313.

BLADON, A (1985), 'Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: a thread', In FALLSIDE, F and WOODS, WA (Eds.), *Computer Speech Processing*, Prentice Hall, New Jersey, Chap. 2.

BODENSTEIN, G and PRAETORIUS, HM (1977), 'Feature extraction from the electroecephalogram by adaptive segmentation', *Proc. IEEE*, Vol. 65, No. 5, pp. 642–652.

BOGERT, BP, HEALY, MJR and TUKEY, JW (1963), 'The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking', In ROSENBLATT, M (Ed.), *Proc. of Symposium on Time Seris Analysis*, John Wiley, New York, pp. 209–243.

BONDER, LJ (1983), 'Between formant space and articulation space', In DEN BROECKE, MPRV and COHEN, A (Eds.), *Proc. 10th Int. Congr. Phonetic Sciences*, Foris, Dordrecht, pp. 347–353.

BOURLARD, H and WELLEKENS, CJ (1989), 'Speech pattern discrimination and multilayer perceptrons', *Computer Speech and Language*, Vol. 3, No. 1, pp. 1–19.

BOVES, L (1984), *The Phonetic Basis of Perceptual Ratings of Running Speech*, Foris, Dordrecht, Holland.

BOYD, I (1987), 'Position reoptimisation for a multipulse excited LPC coder', In LAVER, J and JACK, MA (Eds.), *European Conf. on Speech Technology*, Edinburgh, pp. 37–40.

BRACEWELL, RN (1986), *The Fourier Transform and its Applications*, McGraw-Hill, New York, 2nd ed.

BRIESEMAN, NP (1984), *A New Algorithm For Musical Pitch Estimation*, Master's thesis, University of Canterbury, New Zealand.

BRIESEMAN, NP (19XX), *Computer assistance for the speech impaired*, PhD thesis, University of Canterbury, Christchurch, N.Z.

BRIESEMAN, NP, THORPE, CW and BATES, RHT (1987), 'Nontactile estimation of glottal excitation characteristics of voiced speech', *IEE Proc. Pt. A*, Vol. 134, No. 10, pp. 807–813.

BRIESEMAN, NP, THORPE, CW, ELDER, AG and ROLLS, A (1989), *Sigproc Users Guide and Reference Manual*, Dept. of Elect. Eng., University of Canterbury, Christchurch, N.Z.

BRIGHAM, EO (1974), *The Fast Fourier Transform*, Prentice-Hall, Englewood Cliffs, New Jersey.

BRILL, RL, SEVENICH, ML, SULLIVAN, TJ, SUSTMAN, JD and WITT, RE (1988), 'Behavioral evidence for hearing through the lower jaw by an echolocating dolphin (*Tursiops truncatus*)', *Marine Mammal Science*, Vol. 4, No. 3, pp. 223–230.

BRINCH HANSEN, P (1977), *The Architecture of Concurrent Programs*, Prentice-Hall, Eaglewood Cliffs, N. J.

BROWN, CM (1988), *Human-Computer Interface Design Requirements*, Ablex Publ. Corp., Norwood, N.J.

BURR, DJ (1988), 'Experiments on neural net recognition of spoken and written text', *IEEE Trans. ASSP*, Vol. 36, No. 7, pp. 1162–1168.

BURRUS, CS and PARKS, TW (1985), *DFT/FFT and Convolution Algorithms*, John Wiley, New York.

BURTON, DK (1987), 'Text-independent speaker verification using vector quantization source coding', *IEEE Trans. ASSP*, Vol. 35, No. 2, pp. 133–143.

CANNY, GJ and LEVISON, H (1987), 'The modern management of childhood asthma', *Pediatric Rev. Commun.*, Vol. 1, pp. 123–162.

CARROLL, L (1898), *Through the Looking-Glass*, MacMillan, London.

CARTERETTE, EC and FRIEDMAN, MP (Eds.) (1976), *Handbook of Perception Vol VII: Language and Speech*, Academic Press, New York.

CHABONNEAU, G, RACINEUX, JL, SUDRAUD, M and TUCHAIS, E (1983), 'An accurate recording system and its use in breath sounds spectral analysis', *J. Appl. Physiol.*, Vol. 55, No. 4, pp. 1120–1127.

CHABOT, D (1988), 'A quantitative technique to compare and classify Humpback Whale (*Megaptera novaeangliae*)', *Ethology*, Vol. 77, pp. 89–102.

CHANG, P, GRAY, RM and MAY, J (1987), 'Fourier transform vector quantization for speech coding', *IEEE Trans. Commun.*, Vol. 35, No. 10, pp. 1059–1068.

CHERRY, C (1978), *On Human Communication: A Review, a Survey, and a Criticism*, The MIT Press, Cambridge, Massachusetts, 3rd ed.

CHILDERS, DG (1990), 'Speech processing and synthesis for assessing vocal disorders', *IEEE Eng. Med. Biol. Mag.*, Vol. 9, No. 1, pp. 69–71.

CHILDERS, DG, ALSAKA, YA, HICKS, DM and MOORE, GP (1986), 'Vocal fold vibrations in dysphonia: model vs. measurement', *J. Phon.*, Vol. 14, pp. 429–434.

CHISTOVICH, LA (1968), 'Direction of transition as a perceptual parameter of time-varying stimuli', In *The 6th Int. Cong. on Acoustics*, pp. B99–B102.

CHRISTIANSEN, WN and HÖGBOM, JA (1969), *Radiotelescopes*, Cambridge University Press, Cambridge, UK, 2nd ed.

CLARK, CW (1982), 'The acoustic repertoire of the Southern Right Whale, a quantitative analysis', *Animal Behaviour*, Vol. 30, pp. 1060–1071.

CLARK, CW, MARLER, P and BEEMAN, K (1987), 'Quantitative analysis of animal vocal phonology: An application to swamp sparrow song', *Ethology*, Vol. 76, pp. 101–115.

CLOUTIER, MM (1983), 'The coughing child: Etiology and treatment of a common symptom', *Postgrad. Med.*, Vol. 73, No. 3, pp. 169–175.

COHEN, JR (1989), 'Application of an auditory model to speech recognition', *J. Acoust. Soc. Am.*, Vol. 85, No. 6, pp. 2623–2629.

COHEN, A and LANDSBERG, D (1984), 'Analysis and automatic classification of breath sounds', *IEEE Trans. Biomed. Eng.*, Vol. 31, No. 9, pp. 585–590.

COMREY, AL (1973), *A First Course in Factor Analysis*, Academic Press, New York.

CONNOLLY, N and GODFREY, S (1970), 'Assessment of the child with asthma', *J. Asth. Res.*, Vol. 8, No. 1, pp. 31–36.

CONNOR, RC and NORRIS, KS (1982), 'Are dolphins reciprocal altruists?', *Am. Nat.*, Vol. 119, pp. 358–374.

COOK, CE and BERNFIELD, M (1967), *Radar Signals*, Academic Press, New York.

COOLEY, WW and LOHNES, PR (1971), *Multivariate Data Analysis*, John Wiley, New York.

CORNWELL, TJ (1983), 'A method of stabilizing the clean algorithm', *Astron. Astrophys.*, Vol. 121, pp. 281–285.

CORNWELL, TJ (1988), 'Radio-interferometric imaging of very large objects', *Astron. Astrophys.*, Vol. 202, pp. 316–321.

CORRAO, WM, BROMAN, SS and IRWIN, RS (1979), 'Chronic cough as the sole presenting manifestation of bronchial asthma', *N. Engl. J. Med.*, Vol. 300, pp. 633–637.

COTES, JE (1975), *Lung Function*, Blackwell Scientific Publ., Oxford, 3rd ed.

COX, DR and ISHAM, V (1980), *Point Processes*, Chapman and Hall, London.

CRANEN, B and BOVES, L (1986), 'A parametric voice source model incorporating inter- and intraspeaker variation', In *Speech Input/Output; Techniques and Applications*, IEE, pp. 94–98.

CRANFORD, TW (1988), 'The anatomy of acoustic structures in the spinner dolphin forehead as shown by x-ray computed tomography and computer graphics', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 67–77.

CROCHIERE, RE (1978), 'An analysis of 16kbit/s sub-band coder performance: Dynamic range, tandem connections, and channel errors', *Bell Syst. Tech. J.*, Vol. 57, No. 8, pp. 2927–2952.

CUMMISKEY, P, JAYANT, NS and FLANAGAN, JL (1973), 'Adaptive quantisation in differential PCM coding of speech', *Bell Syst. Tech. J.*, Vol. 52, No. 7, pp. 1105–1118.

CUTTING, JE (1974), 'Two left-hemisphere mechanisms in speech perception', *Perception & Psychophysics*, Vol. 16, No. 3, pp. 601–612.

CUTTING, JE and KAVANAGH, JF (1975), 'On the relationship of speech to language', *J. Am. Speech Hear. Assoc.*, Vol. 17, pp. 500–506.

CUTTING, JE and ROSNER, BS (1974), 'Categories and boundaries in speech and music', *Perception & Psychophysics*, Vol. 16, No. 3, pp. 564–570.

DAINTY, JC (1973), 'Diffraction-limited imaging of stellar objects using telescopes of low optical quality', *Optics Communication*, Vol. 7, No. 2, pp. 129–134.

DAMPER, RI (1982), 'Speech technology — Implications for biomedical engineering', *J. Med. Eng. Tech.*, Vol. 6, No. 4, pp. 135–149.

DARWIN, CJ (1976), 'The perception of speech', In CARTERETTE, EC and FRIEDMAN, MP (Eds.), *Handbook of Perception, Volume VII: Language and Speech*, Academic Press, New York, Chap. 6.

DAUMER, WR (1982), 'Subjective evaluation of several effecient speech coders', *IEEE Trans. Commun.*, Vol. 30, No. 4, pp. 655–662.

DAVEY, BLK (1989), *Advances in Blind Deconvolution*, PhD thesis, University of Canterbury, New Zealand, May.

DAVEY, BLK and THORPE, CW (1987), 'Image and signal reconstruction by shift-and-add', In *IPENZ conference proceedings.*, Inst. Prof. Eng. NZ, Christchurch, May, pp. 147–157.

DAVEY, BLK, COCKE, WJ, BATES, RHT, MCCARTHY, JR., DW, CHRISTOU, JC and COBB, ML (1989), 'Infrared speckle observations of binary Ross 614 AB: Combined shift-and-add and zero-and-add analysis', *Astron. J.*, Vol. 98, No. 3, pp. 1040–1048.

DAVIDSON, G and GERSHO, A (1986), 'Complexity reduction methods for vector excitation coding', In *Int. Conf. on ASSP*, IEEE, Tokyo, pp. 3055–3058.

DAVIS, RO (1986), 'The Personal Acoustics Lab (PAL): a microcomputer-based system for digital signal acquisition, analysis, and synthesis', *Comput. Methods Programs Biomed.*, Vol. 23, pp. 199–210.

DAVIS, SB and MERMELSTEIN, P (1980), 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. ASSP*, Vol. 28, No. 4, pp. 357–366.

DAWSON, SM (1988), 'The high frequency sounds of free-ranging Hector's dolphins, *Cephalorhynchus hectori*', *Rep. Int. Whal. Commn.*, Vol. Special issue 9, pp. 339–344.

DAWSON, SM (1990), *Sounds, Acoustic Behaviour, and Gillnet Entanglement of Hector's Dolphin*, PhD thesis, University of Canterbury, Christchurch, N.Z.

DAWSON, SM (19XXa), 'Incidental catch of Hector's dolphin in inshore gillnets', *J. Wildlife Management*, Submitted for publication.

DAWSON, SM (19XXb), 'Modifying gillnets to reduce entanglement of cetaceans', *J. Wildlife Management*, Submitted for publication.

DAWSON, SV and ELLIOTT, EA (1977), 'Wave-speed limitation on expiratory flow—a unifying concept', *J. Appl. Physiol.*, Vol. 43, pp. 498–515.

DAWSON, SM and JENKINS, PF (1983), 'Chaffinch song repertoires and the Beau Geste hypothesis', *Behaviour*, Vol. 87, pp. 256–269.

DAWSON, SM and SLOOTEN, E (1988), 'Hector's dolphin, *Cephalorhynchus hectori*: Distribution and abundance', *Rep. Int. Whal. Commn.*, Vol. Special Issue 9, pp. 315–324.

DAWSON, SM and THORPE, CW (1990), 'A quantitative analysis of the sounds of Hector's dolphin', *Ethology*, To appear, December 1990.

DE SOUZA, P (1983), 'A statistical approach to the design of an adaptive self-normalizing silence detector', *IEEE Trans. ASSP*, Vol. 31, No. 3, pp. 678–684.

DEBRECZENI, LA, KORPAS, J and SALAT, D (1987), 'Spectral analysis of cough sounds recorded with and without a nose clip', *Bull. Eur. Physiopathol. Respir.*, Vol. 23 (Suppl. 10), pp. 57s–61s.

DEBRECZENI, LA, KORPAS, J, SALAT, D, SADLONOVA-KORPASOVA, J, VÉRTES, C, MASAROVA, E and KAVCOVA, E (1990), 'Spectra of the voluntary first cough sounds', *Acta Physioligica Hungarica*, Vol. 75, No. 2, pp. 117–131.

DÈCINA, M and MODENA, G (1988), 'CCITT standards on digital speech processing', *IEEE J. Sel. Areas Commun.*, Vol. 6, No. 2, pp. 227–234.

DENOEL, E and SOLVAY, J (1985), 'Linear prediction of speech with a least absolute error criterion', *IEEE Trans. ASSP*, Vol. 33, No. 6, pp. 1397–1403.

DIERCKS, KJ, TROCHTA, RT and EVANS, WE (1973), 'Delphinid sonar: measurement and analysis', *J. Acoust. Soc. Am.*, Vol. 54, No. 1, pp. 200–204.

DIMOLITSAS, S (1989), 'Objective speech distortion measures and their relevance to speech quality assessments', *IEE Proc. Pt. I*, Vol. 136, No. 5, pp. 317–324.

DODDINGTON, GR (1985), 'Speaker recognition — Identifying people by their voices', *Proc. IEEE*, Vol. 73, No. 11, pp. 1651–1665.

DORMER, KJ and PHILLIPS, MA (1987), 'Auditory prostheses: Implantable and vibrotactile devices', *IEEE Eng. Med. Biol. Mag.*, Vol. 6, No. 2, pp. 36–41.

DREHER, JJ (1966), 'Cetacean communication: Small group experiment', In NORRIS, KS (Ed.), *Whales, Dolphins and Porpoises*, University of California Press, Berkely, pp. 529–543.

DUDLEY, H (1939), 'Remaking speech', *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177.

DUNCAN, G and JACK, MA (1988), 'Formant estimation algorithm based on pole frequency focussing offering improved noise tolerance and feature resolution', *IEE Proc. Pt. F*, Vol. 135, No. 1, pp. 18–32.

DZIEDZIC, Z (1978), *Experimental Study of the Sonar Emissions of Certain Delphinids, Especially D. Delphis and T. Truncatus*, PhD thesis, University of Paris, France, December. Translated by Gail J. Marciniak-Puckett, K. Jerome Diercks. Applied Research Laboratories, University of Texas, Austin.

DZIEDZIC, A, CHIOLLAZ, M, ESCUDIE, B and HELLION, A (1977), 'On some properties of low frequency sonar signals of the dolphin *Phocoena phocoena*', *Acustica*, Vol. 37, No. 4, pp. 258–266.

EDWARDS, ML and SHRIBERG, LD (1983), *Phonology: Applications in Communicative Disorders*, College-Hill Press, San Diego.

ELDER, AG (19XX), *Aspects of Speaker Recognition*, PhD thesis, University of Canterbury, Christchurch, N.Z.

ELDER, A, BATES, RHT, BRIESEMAN, NP, CLARK, TM, FRIGHT, WR, GARDEN, KL, KENNEDY, WK, SQUIRES, PL, TURNER, SG and THORPE, CW (1987), 'Real-time speech therapy aid', In *Proc. National Electronics Conference*, Auckland, pp. 115–118.

ENDRES, W, BAMBACH, W and FLÖSSER, G (1971), 'Voice spectrograms as a function of age, voice disguise and voice imitation', *J. Acoust. Soc. Am.*, Vol. 49, No. 6 (Pt. 2), pp. 1842–1848.

EVANS, WE (1973), 'Echolocation by marine delphinids and one species of fresh-water dolphin', *J. Acoust. Soc. Am.*, Vol. 54, No. 1, pp. 191–199.

EVANS, WE, AWBREY, FT and HACKBARTH, H (1988), 'High frequency pulses produced by free-ranging Commerson's dolphin (*Cephalorhynchus commersonii*) compared to those of Phocoenids', *Rep. Int. Whal. Commn.*, Vol. Special issue 9, pp. 173–181.

FAIRBANKS, G (1940), *Voice and Articulation Drillbook*, Harper and Row, New York.

FAIRBANKS, G (1958), 'Test of phonemic differentiation: The rhyme test', *J. Acoust. Soc. Am.*, Vol. 30, No. 7, pp. 596–600.

FALLSIDE, F and WOODS, WA (Eds.) (1985), *Computer Speech Processing*, Prentice Hall, New Jersey.

FANT, G (1960), *Acoustic Theory of Speech Production*, Mouton, The Hague.

FANT, G (1968), 'Analysis and synthesis of speech processes', In MALMBERG, B (Ed.), *Manual of Phonetics*, North-Holland, Amsterdam, Chap. 8, pp. 173–277.

FANT, G (1973), *Speech Sounds and Features*, The MIT Press, Cambridge, Massachusetts.

FANT, G (1985), 'Speech technology — Research and development', *Ericsson Review*, Vol. 62, No. 3, pp. 100–107.

FANT, G (1986), 'Glottal flow: models and interaction', *J. Phon.*, Vol. 14, pp. 393–399.

FENTON, TR, PASTERKAMP, H, TAL, A and CHERNICK, V (1985), 'Automated spectral characterization of wheezing in asthmatic children', *IEEE Trans. Biomed. Eng.*, Vol. 32, No. 1, pp. 50–55.

FISSORE, L, LAFACE, P, MICCA, G and PIERACCINI, R (1989), 'Lexical access to large vocabularies for speech recognition', *IEEE Trans. ASSP*, Vol. 37, No. 8, pp. 1197–1213.

FLANAGAN, JL (1972), *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Berlin, 2nd ed.

FLANAGAN, JL (1980), 'Parametric coding of speech spectra', *J. Acoust. Soc. Am.*, Vol. 68, No. 2, pp. 412–419.

FLANAGAN, JL and GOLDEN, RM (1966), 'Phase vocoder', *Bell Syst. Tech. J.*, Vol. 45, pp. 1493–1509.

FLANAGAN, JL, ISHIZAKA, K and SHIPLEY, KL (1975), 'Synthesis of speech from a dynamic model of the vocal cords and vocal tract', *Bell Syst. Tech. J.*, Vol. 54, No. 3, pp. 485–506.

FLANAGAN, JL, ISHIZAKA, K and SHIPLEY, KL (1980), 'Signal models for low bit-rate coding of speech', *J. Acoust. Soc. Am.*, Vol. 68, No. 5, pp. 1535–1555.

FLETCHER, H (1953), *Speech and Hearing in Communication*, D. Van Nostrand, Princeton, NJ.

FORD, JKB and FISHER, HD (1982), 'Killer Whale (*Orcinus orca*) dialects as an indicator of stocks in British Columbia.', *Rep. Int. Whal. Commn.*, Vol. 32, pp. 671–679.

FORGACS, P (1978), *Lung Sounds*, Baillière Tindall, London.

FOWLER, CA (1986), 'An event approach to the study of speech perception from a direct–realist perspective', *J. Phon.*, Vol. 14, No. 1, pp. 3–28.

FRIEDLANDER, B (1982), 'Lattice methods for spectral estimation', *Proc. IEEE*, Vol. 70, No. 9, pp. 990–1017.

FRIED-OKEN, M (1985), 'Voice recognition device as a computer interface for motor and speech impaired people', *Arch. Phys. Med. Rehabil.*, Vol. 66, pp. 678–681.

FRITZELL, B, HAMMARBERG, B, GAUFFIN, J, KARLSSON, I and SUNDBERG, J (1986), 'Breathiness and insufficient vocal fold closure', *J. Phon.*, Vol. 14, pp. 549–553.

FU, KS (1974), *Syntactic Methods in Pattern Recognition*, Academic Press, New York.

FUJIMURA, O (1968), 'An approximation to voice aperiodicity', *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 1, pp. 68–72.

FURUI, S (1981), 'Comparison of speaker recognition methods using statistical features and dynamic features', *IEEE Trans. ASSP*, Vol. 29, No. 3, pp. 342–350.

GABOR, D (1946), 'Theory of communication', *Proc. IEE Pt. III*, Vol. 93, No. 3, pp. 429–441.

GARDENIER, PH, LIM, CA, TAN, DGH and BATES, RHT (1986), 'Aperture distribution phase from single radiation pattern measurement via Gerchberg-Saxton algorithm', *Electronics Letters*, Vol. 22, No. 2, pp. 113–115.

GAUFFIN, J and SUNDBERG, J (1989), 'Spectral correlates of glottal voice source waveform characteristics', *J. Speech Hear. Res.*, Vol. 32, pp. 556–565.

GAVRIELY, N, PALTI, Y and ALROY, G (1981), 'Spectral characteristics of normal breath sounds', *J. Appl. Physiol.*, Vol. 50, No. 2, pp. 307–314.

GAVRIELY, N, PALTI, Y, ALROY, G and GROTBERG, JB (1984), 'Measurement and theory of wheezing breath sounds', *J. Appl. Physiol.*, Vol. 57, No. 2, pp. 481–492.

GHITZA, O (1987), 'Auditory nerve representation criteria for speech analysis/synthesis', *IEEE Trans. ASSP*, Vol. 35, No. 6, pp. 736–740.

GOBL, C (1989), 'A preliminary study of acoustic voice quality correlates', *STL-QPSR*, Vol. 4/1989, pp. 9–22.

GOEDEKING, P (1983), 'A minicomputer-aided method for the detection of features from vocalisations of the cotton-top Tamarin *Saguinus oedipus oedipus*', *Z. Tierpsychol.*, Vol. 62, pp. 321–328.

GOLD, B and RABINER, L (1969), 'Parallel processing techniques for estimating pitch periods of speech in the time domain', *J. Acoust. Soc. Am.*, Vol. 46, No. 2 (Part 2), pp. 442–448.

GOODMAN, DJ and NASH, RD (1982), 'Subjective quality of the same speech transmission conditions in seven different countries', *IEEE Trans. Commun.*, Vol. 30, No. 4, pp. 642–654.

GRAY, RM (1984), 'Vector quantization', *IEEE ASSP Magazine*, Vol. 1, No. 2, pp. 4–29.

GRAY, JR., AH and MARKEL, JD (1974), 'A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis', *IEEE Trans. ASSP*, Vol. 22, No. 3, pp. 207–216.

GRAY, JR., AH and MARKEL, JD (1976), 'Distance measures for speech processing', *IEEE Trans. ASSP*, Vol. 24, No. 5, pp. 380–391.

GREENBERG, S (1988), 'Acoustic transduction in the auditory periphery', *J. Phon.*, Vol. 16, pp. 3–17.

GRIFFIN, DR (1979), 'The early history of research on echolocation', In BUSNEL, R and FISH, JF (Eds.), *Animal Sonar Systems*, Plenum Press, New York, pp. 1–8.

GRIFFIN, DW and LIM, JS (1988), 'Multiband excitation vocoder', *IEEE Trans. ASSP*, Vol. 36, No. 8, pp. 1223–1235.

GROTBERG, JB and DAVIS, SH (1980), 'Fluid-dynamic flapping of a collapsible channel: Sound generation and flow limitation', *J. Biomechanics*, Vol. 13, pp. 219–230.

GUPTA, V, WILSON, TA and BEAVERS, GS (1973), 'A model for vocal cord excitation', *J. Acoust. Soc. Am.*, Vol. 54, No. 6, pp. 1607–1617.

GUTOWSKI, PR, ROBINSON, EA and TREITEL, S (1978), 'Spectral estimation: Fact or fiction', *IEEE Trans. Geosci. Electro.*, Vol. 16, pp. 80–84.

HALLIDAY, D and RESNICK, R (1966), *Physics. Parts I and II*, John Wiley, New York, 2nd ed.

HAMMER, CE and AU, WWL (1980), 'Porpoise echo-recognition: An analysis of controlling target characteristics', *J. Acoust. Soc. Am.*, Vol. 68, No. 5, pp. 1285–1292.

HAMMING, RW (1980), *Coding and Information Theory*, Prentice-Hall, Eaglewood Cliffs, NJ.

HARDCASTLE, WJ (1976), *Physiology of Speech Production*, Academic Press, London.

HARRINGTON, J (1988), 'Acoustic cues for automatic recognition of english consonants', In JACK, MA and LAVER, J (Eds.), *Aspects of Speech Technology*, Edinburgh University Press, Edinburgh, Chap. 2, pp. 69–143.

HARRIS, FJ (1978), 'On the use of windows for harmonic analysis with the discrete Fourier transform', *Proc. IEEE*, Vol. 66, No. 1, pp. 51–83.

HARRIS, CM and WEISS, MR (1963), 'Pitch extraction by computer processing of high-resolution Fourier analysis data', *J. Acoust. Soc. Am.*, Vol. 35, No. 3, pp. 339–343.

HARTLEY, RVL (1928), 'Transmission of information', *Bell Syst. Tech. J.*, Vol. 7, p. 535.

HAWKINS, PR (1973), 'The sound patterns of new zealand english', In *AULLA XV, Proceedings and Papers*, Australasian Universities Language and Literature Association, pp. 13.1–13.8.

HAWKINS, P (1976), 'The role of NZ English in a binary feature analysis of English short vowels', *Jour. of the Intnl. Phonetic Association*, Vol. 6, No. 2, pp. 50–66.

HAYKIN, S (1983), *Communication Systems*, John Wiley, New York, 2nd ed.

HECKER, MHL and GUTTMAN, N (1967), 'Survey of methods for measuring speech quality', *Journal of the Audio Engineering Society*, Vol. 15, No. 4, pp. 400–403.

HENSON, JR., OW and SCHNITZLER, H (1979), 'Performance of airborne biosonar systems: vertebrates other than Microchiroptera', In BUSNEL, R and FISH, JF (Eds.), *Animal Sonar Systems*, Plenum Press, New York, pp. 183–195.

HERMAN, LM and TAVOLGA, WN (1980), 'The communication system of cetaceans', In HERMAN, LM (Ed.), *Cetacean Behavior: Mechanisms and Functions*, John Wiley, New York, Chap. 4, pp. 149–209.

HINCHEY, MJ and SNELLEN, JW (1987), 'Lung sounds and flow limitation', *J. Sound Vib.*, Vol. 116, No. 3, pp. 576–579.

HÖGBOM, JA (1974), 'Aperture synthesis with a non-regular distribution of interferometer baselines', *Astronomical Astrophysics Supplement*, Vol. 15, pp. 417–426.

HOGUE, CWV and FACKRELL, HB (1987), 'The menu workbench : An automatic menu generator for bio-medical programs', *Int. J. Biomed. Comput.*, Vol. 21, pp. 253–264.

HOLMES, JN (1973), 'The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer', *IEEE Trans. Audio Electroacoust.*, Vol. 21, pp. 298–305.

HOUTSMA, AJM and GOLDSTEIN, JL (1972), 'The central origin of the pitch of complex tones: Evidence from musical interval recognition', *J. Acoust. Soc. Am.*, Vol. 51, No. 2 (Pt. 2), pp. 520–529.

HUGGINS, AWF and NICKERSON, RS (1985), 'Speech quality evaluation using "phoneme-specific" sentences', *J. Acoust. Soc. Am.*, Vol. 77, No. 5, pp. 1896–1906.

HUGHES, GW and HALLE, M (1956), 'Spectral properties of fricative consonants', *J. Acoust. Soc. Am.*, Vol. 28, No. 2, pp. 303–310.

HUNT, BR, FRIGHT, WR and BATES, RHT (1983), 'Analysis of the shift-and-add method for imaging through turbulent media', *J. Opt. Soc. Am.*, Vol. 73, No. 4, pp. 456–465.

IEEE (1969), 'Recommended practice for speech quality measurements', *IEEE Trans. Audio Electroacoust.*, Vol. 17, No. 3, pp. 225–246.

IMAIZUMI, S (1986), 'Acoustic measures of roughness in pathological voice', *J. Phon.*, Vol. 14, pp. 457–462.

IMMENDÖRFER, M (1986), 'Applications for speech processing in telecommunication and office equipment.', *Electrical Communications*, Vol. 60, No. 1, pp. 71–78.

INMON, WH (1981), *Effective Database Design*, Prentice-Hall, Eaglewood Cliffs, N. J.

IRWIN, RS, ROSEN, MJ and BRAMAN, SS (1977), 'Cough: A comprehensive review', *Arch. Intern. Med.*, Vol. 137, pp. 1186–1191.

IRWIN, RS, CURLEY, FJ and FRENCH, CL (1990), 'Chronic cough', *Am. Rev. Respir. Dis.*, Vol. 141, pp. 640–647.

ISHIZAKA, K (1981), 'Equivalent lumped-mass models of vocal fold vibration', In STEVENS, KN and HIRANO, M (Eds.), *Vocal Fold Physiology*, University of Tokyo Press, Tokyo, Chap. 17, pp. 231–244.

ISHIZAKA, K and FLANAGAN, JL (1972), 'Synthesis of voiced sounds from a two-mass model of the vocal cords', *Bell Syst. Tech. J.*, Vol. 51, No. 6, pp. 1233–1268.

ITAKURA, F and SAITO, S (1968), 'Analysis synthesis telephony based in the maximum likelihood method', In *The 6th Int. Cong. on Acoustics*, Tokyo, pp. C17–C20.

ITAKURA, F and SAITO, S (1970), 'A statistical method for estimation of speech spectral density and formant frequencies', *Electron. Commun. Japan*, Vol. 53-A, No. 1, pp. 36–43.

JAKATDAR, P and MULLA, HD (1986), 'Speech communication for personal computers', *Electrical Communications*, Vol. 60, No. 1, pp. 79–86.

JAKOBSON, R (1978), *Six Lectures on Sound and Meaning*, The Harvester Press, Hassocks, England. Translated from the French by John Mepham.

JAKOBSON, R, FANT, CGM and HALLE, M (1961), *Preliminaries to Speech Analysis: The distinctive features and their correlates*, MIT Press, Cambridge, Mass.

JAYANT, NS (1974), 'Digital coding of speech waveforms: PCM, DPCM, and DM quantizers', *Proc. IEEE*, Vol. 62, No. 5, pp. 611–632.

JAYANT, NS (1990), 'High-quality coding of telephone speech and wideband audio', *IEEE Trans. Commun.*, Vol. 28, No. 1, pp. 10–20.

JAYANT, NS and NOLL, P (1984), *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, Eaglewood Cliffs, NJ.

JENSEN *et al.* (1988), *TopSpeed Modula-2 Users Manual*, Jensen & Partners International, Inc., 1101 San Antonio Rd., Suite 301, Mountain View, CA.

JOHNS-LEWIS, CM (1986), 'Digital analysis of pitch and silence in three speech styles', In *Int. Conf. on Speech Input/Output; techniques and applications*, IEE, pp. 281–286.

JONES, RS (1966), 'Assessment of respiratory function in the asthmatic child', *Br. Med. J.*, Vol. 2, pp. 972–975.

JOSEPHS, LK, GREGG, I and HOLGATE, ST (1990), 'Does non-specific bronchial responsiveness indicate the severity of asthma?', *Eur. Respir. J.*, Vol. 3, pp. 220–227.

JUANG, B, RABINER, LR and WILPON, JG (1987), 'On the use of bandpass liftering in speech recognition', *IEEE Trans. ASSP*, Vol. 35, No. 7, pp. 947–954.

KAMMINGA, C (1988), 'Echolocation signal types of odontocetes', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 9–22.

KAMMINGA, C and WIERSMA, H (1981), 'Investigations on cetacean sonar II. Acoustical similarities and differences in odontocete sonar signals', *Aquatic Mammals*, Vol. 8, No. 2, pp. 41–62.

KAMMINGA, C and WIERSMA, H (1982), 'Investigations on cetacean sonar V. The true nature of the sonar sound of *Cephalorynchus commersonii*', *Aquatic Mammals*, Vol. 9, No. 3, pp. 95–104.

KANG, GS and EVERETT, SS (1985), 'Improvement of the excitation source in the narrow-band linear prediction vocoder', *IEEE Trans. ASSP*, Vol. 33, No. 2, pp. 377–386.

KARLSSON, I (1986), 'Glottal wave forms for normal female speakers', *J. Phon.*, Vol. 14, pp. 415–419.

KASSLING, B (1989), 'Experiences from two simple tactile aids as support during speechreading', *STL-QPSR*, No. 1/1989, pp. 155–157.

KASUYA, H, OGAWA, S, KIKUCHI, Y and EBIHARA, S (1986), 'An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology', *Speech Communication*, Vol. 5, No. 2, pp. 171–181.

KAUFMAN, RE (1978), *A Fortran Coloring Book*, MIT Press, Cambridge, Mass.

KAY, SM and MARPLE, JR., SL (1981), 'Spectrum analysis—A modern perspective', *Proc. IEEE*, Vol. 69, No. 11, pp. 1380–1419.

KELEMEN, SA, CSERI, T and MAROZSAN, I (1987), 'Information obtained from tussigrams and the possibilities of their application in medical practice', *Bull. Eur. Physiopathol. Respir.*, Vol. 23 (Suppl. 10), pp. 51s–56s.

KERSTA, LG (1962), 'Voiceprint identification', *Nature*, Vol. 196, No. 4861, pp. 1253–1257.

KINSLER, LE, FREY, AR, COPPENS, AB and SANDERS, JV (1982), *Fundamentals of Acoustics*, John Wiley, New York, 3rd ed.

KITAWAKI, N and NAGABUCHI, H (1988), 'Quality assessment of speech coding and speech synthesis systems', *IEEE Commun. Mag.*, Vol. 26, No. 10, pp. 36–44.

KITAWAKI, N, NAGABUCHI, H and ITOH, K (1988), 'Objective quality evaluation for low-bit-rate speech coding systems', *IEEE J. Sel. Areas Commun.*, Vol. 6, No. 2, pp. 242–247.

KLATT, DH (1987), 'Review of test-to-speech conversion for English', *J. Acoust. Soc. Am.*, Vol. 82, No. 3, pp. 737–793.

KNORR, SG (1979), 'Reliable voiced/unvoiced decision', *IEEE Trans. ASSP*, Vol. 27, No. 3, pp. 263–267.

KOENIG, W, DUNN, HK and LACY, LY (1946), 'The sound spectrograph', *J. Acoust. Soc. Am.*, Vol. 18, No. 1, pp. 19–49.

KOHONEN, T (1988), 'The "neural" phonetic typewriter', *Computer*, Vol. 21, No. 3, pp. 11–22.

KOIZUMI, T, TANIGUCHI, S and HIROMITSU, S (1985), 'Glottal source–vocal tract interaction', *J. Acoust. Soc. Am.*, Vol. 78, No. 5, pp. 1541–1547.

KOIZUMI, T, TANIGUCHI, S and HIROMITSU, S (1987), 'Two-mass models of the vocal cords for natural sounding voice synthesis', *J. Acoust. Soc. Am.*, Vol. 82, No. 4, pp. 1179–1192.

KOLSTON, PJ (1989), *Towards a Better Understanding of Cochlear Mechanics: A New Cochlear Model*, PhD thesis, University of Canterbury, Christchurch, N.Z.

KOPP, GA and GREEN, HC (1946), 'Phonetic principles of visible speech', *J. Acoust. Soc. Am.*, Vol. 18, No. 1, pp. 74–89.

KORPAS, J, SADLONOVA, J, SALAT, D and MASAROVA, E (1987), 'The origin of cough sounds', *Bull. Eur. Physiopathol. Respir.*, Vol. 23 (Suppl. 10), pp. 47s–50s.

KREYSZIG, E (1970), *Introductory Mathematical Statistics*, John Wiley, New York.

KREYSZIG, E (1979), *Advanced Engineering Mathematics*, John Wiley, New York, 5th ed.

KRISHNAIAH, PR and KANAL, LN (Eds.) (1982), *Classification, Pattern Recognition and Reduction of Dimensionality*, North-Holland, Amsterdam.

KRISHNAMURTHY, AK and CHILDERS, DG (1986), 'Two-channel speech analysis', *IEEE Trans. ASSP*, Vol. 34, No. 4, pp. 730–743.

KROON, P and DEPRETTERE, EF (1988), 'A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16kbit/s', *IEEE J. Sel. Areas Commun.*, Vol. 6, No. 2, pp. 353–363.

KROON, P, DEPRETTERE, EF and SLUYTER, RJ (1986), 'Regular-pulse excitation—A novel approach to effective and efficient multipulse coding of speech', *IEEE Trans. ASSP*, Vol. 34, No. 5, pp. 1054–1063.

LADEFOGAD, P (1982), *A Course in Phonetics*, Harcourt Brace Jovanovich, New York, 2nd ed.

LANE, RG and BATES, RHT (1987), 'Automatic multi-dimensional deconvolution', *J. Opt. Soc. Am. A*, Vol. 4, No. 1, pp. 180–188.

LAVER, J (1980), *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge.

LEE, C (1988), 'On robust linear prediction of speech', *IEEE Trans. ASSP*, Vol. 36, No. 5, pp. 642–650.

LEFEVRE, J and PASSIEN, O (1985), 'Efficient algorithms for obtaining multipulse excitation for LPC coders', In *Int. Conf. on ASSP*, IEEE, pp. 957–960.

LEVITT, H (1973), 'Speech processing aids for the deaf: an overview', *IEEE Trans. Audio Electroacoust.*, Vol. 21, No. 3, pp. 269–273.

LIEBERMAN, P (1963), 'Some acoustic measures of the fundamental periodicity of normal and pathologic larynges', *J. Acoust. Soc. Am.*, Vol. 35, No. 3, pp. 344–353.

LIEBERMAN, P (1975), *On the Origins of Language*, Macmillan, New York.

LIEBERMAN, P and BLUMSTEIN, SE (1988), *Speech Physiology, Speech Perception, and Acoustic Phonetics*, Cambridge University Press, Cambridge.

LILLY, JC and MILLER, AM (1961), 'Sounds emitted by the bottlenose dolphin', *Science*, Vol. 133, pp. 1689–1692.

LINGGARD, R (1985), *Electronic Synthesis of Speech*, Cambridge University Press, Cambridge.

LIPPMANN, RP (1987), 'An introduction to computing with neural nets', *IEEE ASSP Magazine*, Vol. 4, No. 2, pp. 4–22.

LÖFQVIST, A (1986), 'The long-time-average spectrum as a tool in voice research', *J. Phon.*, Vol. 14, pp. 471–475.

LOUDON, R and MURPHY, JR., RLH (1984), 'Lung sounds', *Am. Rev. Respir. Dis.*, Vol. 130, pp. 663–673.

LOUDON, RG and SHAW, GB (1967), 'Mechanics of cough in normal subjects and in patients with obstructive respiratory disease', *Am. Rev. Respir. Dis.*, Vol. 96, pp. 666–677.

LYON, RF and MEAD, C (1988), 'An analog electronic cochlear', *IEEE Trans. ASSP*, Vol. 36, No. 7, pp. 1119–1134.

MACKLEM, PT (1974), 'Physiology of cough', *Ann. Otol.*, Vol. 83, pp. 761–768.

MACLAGAN, MA (1982), 'An acoustic study of New Zealand vowels', *N.Z. Speech Therapist's Journal*, Vol. 37, No. 1, pp. 20–26.

MAEDA, S, FUJIMURA, O and SAWASHIMA, M (1968), 'Voice aperiodicity in terms of pulse shape and interval', In *The 6th Int. Cong. on Acoustics*, Tokyo, pp. B43–B46.

MAKHOUL, J (1973), 'Spectral analysis of speech by linear prediction', *IEEE Trans. Audio Electroacoust.*, Vol. 21, No. 3, pp. 140–148.

MAKHOUL, J (1975), 'Linear prediction: A tutorial review', *Proc. IEEE*, Vol. 63, No. 4, pp. 561–580.

MAKHOUL, JI and WOLF, JJ (1972), *Linear Prediction and the Spectral Analysis of Speech*, Technical Report BBN-2304, Bolt Beranek and Newman Inc., Cambridge, Mass., 31 August.

MANGOLD, H (1988), 'Principles of automatic processing of speech signals and their application in medical technology and for aids for handicapped', *Medical Progress through Technology*, Vol. 14, pp. 39–56.

MARK, JW and TODD, TD (1981), 'A nonuniform sampling approach to data compression', *IEEE Trans. Commun.*, Vol. 29, No. 1, pp. 24–32.

MARKEL, JD (1972), 'The SIFT algorithm for fundamental frequency estimation', *IEEE Trans. Audio Electroacoust.*, Vol. 20, No. 5, pp. 367–377.

MARKEL, JD and GRAY, JR., AH (1976), *Linear Prediction of Speech*, Springer-Verlag, Berlin.

MARKEL, JD, OSHIKA, BT and GRAY, JR., AH (1977), 'Long-term feature averaging for speaker recognition', *IEEE Trans. ASSP*, Vol. 25, No. 4, pp. 330–337.

MARLER, P and PETERS, S (1981), 'Sparrows learn adult song and more from memory.', *Science*, Vol. 213, pp. 780–782.

MATHEWS, MV (1959), 'Extremal coding for speech transmission', *IRE Trans. Inf. Theory*, Vol. 5, No. 3, pp. 129–136.

MCADAM, DW and WHITAKER, HA (1971), 'Language production: Electroencephalographic localization in the normal human brain', *Science*, Vol. 172, No. 3982, pp. 499–502.

MCAULAY, RJ and QUATIERI, TF (1986), 'Speech analysis/synthesis based on a sinusoidal representation', *IEEE Trans. ASSP*, Vol. 34, No. 4, pp. 744–754.

MCBRIDE, AF (1956), 'Evidence for echolocation by cetaceans', *Deep Sea Research*, Vol. 3, pp. 153–154.

MCCANDLESS, SS (1974), 'An algorithm for automatic formant extraction using linear prediction spectra', *IEEE Trans. ASSP*, Vol. 22, No. 2, pp. 135–141.

MCFADDEN, ER (1984), 'Exercise performance in the asthmatic', *Euro. Rev. Respir. Dis.*, Vol. 129:Suppl., pp. S84–S87.

MERMELSTEIN, P (1973), 'Articulatory model for the study of speech production', *J. Acoust. Soc. Am.*, Vol. 53, No. 4, pp. 1070–1082.

MERMELSTEIN, P (1979), 'Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech', *J. Acoust. Soc. Am.*, Vol. 66, No. 6, pp. 1664–1667.

METAGRAPHICS (1986), *MetaWindow Reference Manual*, MetaGraphic Software Corporation, 4575 Scotts Valley Drive, CA.

MILENKOVIC, P (1986), 'Glottal inverse filtering by joint estimation of an AR system with a linear input model', *IEEE Trans. ASSP*, Vol. 34, No. 1, pp. 28–42.

MILLANE, RP and BATES, RHT (1982), 'Inverse methods for branched ducts and transmission lines', *IEE Proc. Pt. F*, Vol. 129, No. 1, pp. 45–51, See also BATES, R.H.T. and MILLANE, R.P. (1981), 'Time domain approach to inverse scattering', *IEEE Trans. Ant. Prop.*, Vol. 29, No. 2, pp. 359–363.

MILLER, RL (1959), 'Nature of the vocal cord wave', *J. Acoust. Soc. Am.*, Vol. 31, No. 6, pp. 667–677.

MILLER, EH (1979), 'An approach to the analysis of graded vocalisations of birds', *Behav. Neural Biol.*, Vol. 27, pp. 25–38.

MOLINGER, D (1975), *Aspects of Language*, Harcourt Brace Jovanovich, New York, 2nd ed.

MORGANE, PJ, JACOBS, MS and GALABURDA, A (1986), 'Evolutionary morphology of the dolphin brain.', In SCHUSTERMAN, RJ, THOMAS, JA and WOOD, FG (Eds.), *Dolphin Cognition and Behaviour: A comparative Approach*, Lawrence Erlbaum Associates, New Jersey, pp. 5–29.

MORRIS, D (1985), 'Phase retrieval in the radio holography of reflector antennas and radio telescopes', *IEEE Trans. Antenn. Prop.*, Vol. 33, No. 7, pp. 749–755, See also correction in Vol. AP-33, p. 1419.

MORSE, PM and INGARD, KU (1968), *Theoretical Acoustics*, McGraw-Hill, New York.

MURCHISON, AE (1979), 'Detection range and range resolution of echolocating Bottlenose porpoise (*Tursiops truncatus*)', In BUSNEL, R and FISH, JF (Eds.), *Animal Sonar Systems*, Plenum Press, New York, pp. 43–70.

NAPIER, PJ, THOMPSON, AR and EKERS, RD (1983), 'The Very Large Array: Design and performance of a modern synthesis radio telescope', *Proc. IEEE*, Vol. 71, No. 11, pp. 1295–1320.

NARAYAN, R and NITYANANDA, R (1986), 'Maximum entropy image restoration in astronomy', *Ann. Rev. Astron. Astrophys.*, Vol. 24, pp. 127–170.

NATVIG, JE (1988), 'Evaluation of six medium bit-rate coders for the Pan-European digital mobile radio system', *IEEE J. Sel. Areas Commun.*, Vol. 6, No. 2, pp. 324–331.

NEUGEBAUER, O (1957), *The Exact Sciences in Antiquity*, Brown University Press, Providence, 2nd ed.

NOCERINO, N, SOONG, FK, RABINER, LR and KLATT, DH (1985), 'Comparative study of several distortion measures for speech recognition', *Speech Communication*, Vol. 4, pp. 317–331.

NOLL, P (1975), 'A comparative study of various quantization schemes for speech encoding', *Bell Syst. Tech. J.*, Vol. 54, No. 9, pp. 1597–1614.

NORRIS, KS (1969), 'The echolocation of marine mammals', In ANDERSEN, HT (Ed.), *The Biology of Marine Mammals*, Academic Press, New York, Chap. 10, pp. 391–423.

NORRIS, KS and EVANS, WE (1966), 'Directionality of echolocation clicks in the rough-tooth porpoise, *Steno bredanensis* (Lesson)', In TAVOLGA, WN (Ed.), *Marine Bio-acoustics*, Pergamon, Oxford, pp. 305–316.

NYQUIST, H (1924), 'Certain factors affecting telegraph speed', *Bell Syst. Tech. J.*, Vol. 3, p. 324.

OPPENHEIM, AV (1969), 'Speech analysis-synthesis system based on homomorphic filtering', *J. Acoust. Soc. Am.*, Vol. 45, No. 2, pp. 458–465.

OPPENHEIM, AV and SCHAFER, RW (1968), 'Homomorphic analysis of speech', *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 2, pp. 221–226.

OPPENHEIM, AV and SCHAFER, RW (1975), *Digital Signal Processing*, Prentice-Hall, New Jersey.

OPPENHEIM, AV and WILLSKY, AS (1983), *Signals and Systems*, Prentice-Hall, New Jersey. (Ian T. Young, associate author).

OPPENHEIM, AV, SCHAFER, RW and STOCKHAM, TG (1968), 'Nonlinear filtering of multiplied and convolved signals', *Proc. IEEE*, Vol. 56, No. 8, pp. 1264–1991.

O'SHAUGHNESSY, D (1986), 'Speaker recognition', *IEEE ASSP Magazine*, Vol. 3, No. 4, pp. 4–17.

PALIWAL, KK and RAO, PVS (1981), 'A modified autocorrelation method of linear prediction for pitch-synchronous analysis of voiced speech', *Signal Processing*, Vol. 3, No. 2, pp. 181–185.

PAPOULIS, A (1984), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, Auckland, 2nd ed.

PASTERKAMP, H, CARSON, C, DAIEN, D and OH, Y (1989), 'Digital respirosonography: New images of lung sounds', *Chest*, Vol. 96, No. 6, pp. 1405–1412.

PAYNE, R and WEBB, D (1971), 'Orientation by means of long range acoustic signalling in baleen whales', *Annals New York Academy of Sciences*, Vol. 188, pp. 110–141.

PERKELL, JS and COHEN, MH (1989), 'An indirect test of the quantal nature of speech in the production of the vowels /i/, /a/ and /u/', *J. Phon.*, Vol. 17, No. 1/2, pp. 123–133.

PETERSON, GE (1952), 'The information-bearing elements of speech', *J. Acoust. Soc. Am.*, Vol. 24, No. 6, pp. 629–637.

PICKETT, JM (1972), 'Status of speech analyzing communication aids for the deaf', *IEEE Trans. Audio Electroacoust.*, Vol. 20, No. 1, pp. 3–8.

PIIRILÄ, P and SOVIJÄRVI, ARA (1989), 'Differences in acoustic and dynamic characteristics of spontaneous cough in pulmonary diseases', *Chest*, Vol. 96, No. 1, pp. 46–53.

POPPER, AN (1980), 'Sound emission and detection by delphinids', In HERMAN, LM (Ed.), *Cetacean Behavior: Mechanisms and Functions*, John Wiley, New York, Chap. 1, pp. 1–52.

RABINER, LR and JUANG, BH (1986), 'An introduction to hidden Markov models', *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4–16.

RABINER, LR and LEVINSON, SE (1981), 'Isolated and connected word recognition—Theory and selected applications', *IEEE Trans. Commun.*, Vol. 29, No. 5, pp. 621–659.

RABINER, LR and SCHAFER, RW (1978), *Digital Signal Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, New Jersey 07632, USA.

RABINER, LR, SAMBUR, MR and SCHMIDT, CE (1975), 'Applications of a nonlinear smoothing algorithm to speech processing', *IEEE Trans. ASSP*, Vol. 23, No. 6, pp. 552–557.

RABINER, LR, ATAL, BS and SAMBUR, MR (1977), 'LPC prediction error—analysis of its variation with the position of the analysis frame', *IEEE Trans. ASSP*, Vol. 25, No. 5, pp. 434–442.

RABINER, LR, WILPON, JG and SOONG, FK (1989), 'High performance connected digit recognition using hidden Markov models', *IEEE Trans. ASSP*, Vol. 37, No. 8, pp. 1214–1225.

RAMACHANDRAN, RP and KABAL, P (1989), 'Pitch prediction filters in speech coding', *IEEE Trans. ASSP*, Vol. 37, No. 4, pp. 467–478.

REMEZ, RE (1986), 'Realism, language, and another barrier', *J. Phon.*, Vol. 14, No. 1, pp. 89–97.

REMEZ, RE, RUBIN, PE, PISONI, DB and CARRELL, TD (1981), 'Speech perception without traditional speech cues', *Science*, Vol. 212, pp. 947–950.

REYNOLDS, AJ (1974), *Turbulent Flows in Engineering*, John Wiley, London.

RICHARDS, DG (1985), 'Biological strategies for communication', *IEEE Commun. Mag.*, Vol. 23, No. 6, pp. 10–18.

RIDGWAY, SH and CARDER, DA (1988), 'Nasal pressure and sound production in an echolocating White Whale *Delphinapterus leucas*', In NACHTIGALL, PE and MOORE, PWB (Eds.), *Animal Sonar, Processes and Performance*, Plenum Press, New York, pp. 53–60.

VAN RIPER, C and IRWIN, JV (1958), *Voice and Articulation*, Prentice Hall, New Jersey.

ROBERTS, RA and MULLIS, CT (1987), *Digital Signal Processing*, Addison-Wesley, Reading, Mass.

RODDIER, F (1988), 'Interferometric imaging in optical astronomy', *Physics Reports*, Vol. 170, No. 2, pp. 97–166.

RORABAUGH, B (1986), *Signal Processing Design Techniques*, TAB Professional and Reference Books, PA, USA.

ROSENBERG, AE (1971), 'Effect of glottal pulse shape on the quality of natural vowels', *J. Acoust. Soc. Am.*, Vol. 49, No. 2 (Pt. 2), pp. 583–590.

ROSS, BB, GRAMIAK, R and RAHN, H (1955), 'Physical dynamics of the cough mechanism', *J. Appl. Physiol.*, Vol. 8, pp. 264–268.

ROTHAUSER, EH, URBANEK, GE and PACHL, WP (1968), 'Isopreference method for speech evaluation', *J. Acoust. Soc. Am.*, Vol. 44, No. 2, pp. 408–418.

ROTHAUSER, EH, URBANEK, GE and PACHL, WP (1971), 'A comparison of preference measurement methods', *J. Acoust. Soc. Am.*, Vol. 49, No. 4 (Pt. 2), pp. 1297–1308.

ROTHENBERG, M (1973), 'A new inverse-filtering technique for deriving the glottal air flow waveform during voicing', *J. Acoust. Soc. Am.*, Vol. 53, No. 6, pp. 1632–1645.

ROTHENBERG, M (1981), 'Acoustic interaction between the glottal source and the vocal tract', In STEVENS, KN and HIRANO, M (Eds.), *Vocal Fold Physiology*, University of Tokyo Press, Tokyo, Chap. 21, pp. 305–328.

ROTHENBERG, M (1983), 'An interactive model for the voice source', In BLESS, DM and ABBS, JH (Eds.), *Vocal Fold Physiology*, College-Hill Press, San Diego, Chap. 12, pp. 155–165.

RUBOW, R (1984), 'Role of feedback, reinforcement, and compliance on training and transfer in biofeedback-based rehabilitation of motor speech disorders', In MCNEIL, MR, ROSENBEK, JC and ARONSON, AE (Eds.), *The Dysarthrias: Physiology, Acoustics, Perception, Management*, College-Hill, San Diego, Chap. 8, pp. 207–230.

SALÁT, D, KORPÁŠ, J, SALÁTOVÁ, V, KORPÁŠOVA-SADLOŇOVÁ, J and PALEČEK, D (1987), 'The Tussiphonogram during asthmatic attack', *Acta Physiol. Hung.*, Vol. 70, No. 2–3, pp. 223–225.

DE SAUSSURE, F (1959), *Course in General Linguistics*, McGraw-Hill, New York. Originally publ. 1916, Translated from the French by Wade Baskin.

SAVOJI, MH (1989), 'A robust algorithm for accurate endpointing of speech signals', *Speech Communication*, Vol. 8, No. 1, pp. 45–60.

SCARR, RWA (1968), 'Zero crossings as a means of obtaining spectral information in speech analysis', *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 2, pp. 247–255.

SCHNITZLER, H (1968), 'Die ultraschall-ortungslaute der hufeisen-feldermäuse (*Chiroptera Rhinolophidae*) in verschiedenen orientierungssituationen', *Z. vergl. Physiol.*, Vol. 57, pp. 376–408.

SCHROEDER, MR (1966), 'Vocoders: Analysis and synthesis of speech', *Proc. IEEE*, Vol. 54, No. 5, pp. 720–734.

SCHROEDER, MR (1968), 'Reference signal for signal quality studies', *J. Acoust. Soc. Am.*, Vol. 44, No. 6, pp. 1735–1736.

SCHROEDER, MR (1975), 'Models of hearing', *Proc. IEEE*, Vol. 63, No. 9, pp. 1332–1350.

SCHROEDER, MR (1983), 'Speech and hearing: Some important interactions', In DEN BROECKE, MPRV and COHEN, A (Eds.), *Proc. 10th Int. Congr. Phonetic Sciences*, Foris, Dordrecht, pp. 41–52.

SCHROEDER, MR (1984), 'Linear prediction, entropy and signal analysis', *IEEE ASSP Magazine*, Vol. 1, No. 3, pp. 3–11.

SCHROEDER, MR, ATAL, BS and HALL, JL (1979), 'Objective measure of certain speech signal degradations based on masking properties of human auditory perception', In LINDBLOM, B and ÖHMAN, S (Eds.), *Frontiers of Speech Communication Research*, Academic Press, London, pp. 217–229. (Also in *J. Acoust Soc. Am.*, Vol. 66, No. 6, pp. 1647–1652, 1979.).

SCHWARTZ, UJ (1978), 'Mathematical-statistical description of the iterative beam removing technique (Method CLEAN)', *Astron. Astrophys.*, Vol. 65, pp. 345–356.

SCHWARTZ, M and SHAW, L (1975), *Signal Processing*, MgGraw-Hill, New York.

SCOTT, S and CAIRD, FI (1983), 'Speech therapy for Parkinson's disease', *J. of Neurology, Neurosurgery, and Psychiatry*, Vol. 46, pp. 140–144.

SEYFARTH, RM, CHENEY, DL and MARLER, P (1980a), 'Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication', *Science*, Vol. 210, pp. 801–803.

SEYFARTH, RM, CHENEY, DL and MARLER, P (1980b), 'Vervet monkey alarm calls: Semantic communication in a free-ranging primate', *Anim. Behav.*, Vol. 28, pp. 1070–1094.

SHANNON, CE (1948), 'A mathematical theory of communication', *Bell Syst. Tech. J.*, Vol. 27, No. 3,4, pp. 379–423, 623–656.

SHANNON, CE (1949), 'Communication in the presence of noise', *Proc. of the IRE*, Vol. 37, No. 1, pp. 10–21.

SIEGEL, LJ (1979), 'A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier', *IEEE Trans. ASSP*, Vol. 27, No. 1, pp. 83–89.

SIEGEL, LJ and BESSEY, AC (1982), 'Voiced/unvoiced/mixed excitation classification of speech', *IEEE Trans. ASSP*, Vol. 30, No. 3, p. 451.

SINGHAL, S and ATAL, BS (1983), 'Optimising LPC filter parameters for multipulse excitation', In *Int. Conf. on ASSP*, IEEE, pp. 781–784.

SINGHAL, S and ATAL, BS (1989), 'Amplitude optimization and pitch prediction in multipulse coders', *IEEE Trans. ASSP*, Vol. 37, No. 3, pp. 317–327.

SJARE, BL and SMITH, TG (1986), 'The relationship between behavioral activity and underwater vocalizations of the white whale, *Delphinapterus leucas*', *Can. J. Zool.*, Vol. 64, pp. 2824–2831.

SKINNER, PH and SHELTON, RL (Eds.) (1978), *Speech, Language, and Hearing: Normal Processes and Disorders*, Addison-Wesley, Reading, Mass.

SKOLNIK, MI (Ed.) (1980), *Introduction to Radar Systems*, McGraw Hill Kogakusha, Tokyo, 2nd ed.

SLEPIAN, D (1976), 'On bandwidth', *Proc. IEEE*, Vol. 64, No. 3, pp. 292–300.

SLOOTEN, E and DAWSON, SM (1988), 'Studies on Hector's dolphin, *Cephalorynchus hectori*: a progress report', *Rep. Int. Whal. Commn.*, Vol. Special Issue 9, pp. 325–338.

SNEDECOR, GW and COCHRAN, WG (1980), *Statistical Methods*, Iowa State University Press, Iowa, 7th ed.

SONDHI, MM (1968), 'New methods of pitch extraction', *IEEE Trans. Audio Electroacoust.*, Vol. 16, No. 2, pp. 262–266.

SONDHI, MM (1974), 'Model for wave propogation in a lossy vocal tract', *J. Acoust. Soc. Am.*, Vol. 55, No. 5, pp. 1070–1075.

SONDHI, MM (1979), 'Estimation of vocal-tract areas: The need for acoustic measurements', *IEEE Trans. ASSP*, Vol. 27, No. 3, pp. 268–273.

SONDHI, MM (1984), 'A survey of the vocal tract inverse problem: Theory, computation and experiments', In SANTOSA, F, PAO, Y, SYMES, WW and HOLLAND, C (Eds.), *Inverse Problems of Acoustic and Elasitc Waves*, SIAM, Philadelphia.

SONDHI, MM and GOPINATH, B (1971), 'Determination of vocal-tract shape from impulse response at the lips', *J. Acoust. Soc. Am.*, Vol. 49, No. 6 (Pt. 2), pp. 1867–1873.

SONDHI, MM and SCHROETER, J (1987), 'A hybrid time-frequency domain articulatory speech synthesizer', *IEEE Trans. ASSP*, Vol. 35, No. 7, pp. 955–967.

SOONG, FK, ROSENBERG, AE, RABINER, LR and JUANG, BH (1985), 'A vector quantization approach to speaker recognition', In *Int. Conf. on ASSP*, IEEE, Tampa, Fla., pp. 387–390.

SPARLING, DW and WILLIAMS, JD (1978), 'Multivariate analysis of avian vocalisations', *J. Theor. Biol.*, Vol. 74, pp. 83–107.

SREENIVAS, TV (1988), 'Modelling LPC-residue by components for good quality speech coding', In *Int. Conf. on ASSP*, pp. 171–174.

STARMER, CF, CHERVENY, DJ, DIETZ, MA and SMALTZ, JM (1987), 'Managing research data with self-documenting files', *Comput. Biomed. Res.*, Vol. 20, pp. 264–278.

STEVENS, KN (1985), 'Spectral prominences and phonetic distinctions in language', *Speech Communication*, Vol. 4, pp. 137–144.

STEVENS, KN (1989), 'On the quantal nature of speech', *J. Phon.*, Vol. 17, pp. 3–45.

STEVENS, KN and HOUSE, AS (1955), 'Development of a quantitative description of vowel articulation', *J. Acoust. Soc. Am.*, Vol. 27, pp. 484–493.

STEVENS, KN and HOUSE, AS (1972), 'Speech perception', In TOBIAS, JV (Ed.), *Foundations of Modern Auditory Theory*, Academic Press, New York, Chap. 1.

STOCKHAM, JR., TG, CANNON, TM and INGEBRETSEN, RB (1975), 'Blind deconvolution through digital signal processing', *Proc. IEEE*, Vol. 63, No. 4, pp. 678–692.

STREMLER, FG (1982), *Introduction to Communication Systems*, Addison-Wesley, Reading, Massachusetts, 2nd ed.

STROHBEHN, JW (1968), 'Line-of-sight wave propagation through the turbulent atmosphere', *Proc. IEEE*, Vol. 56, No. 8, pp. 1301–1318.

SUNDBERG, J and GAUFFIN, J (1979), 'Waveform and spectrum of the glottal voice source', In LINDBLOM, B and ÖHMAN, S (Eds.), *Frontiers of Speech Communication Research*, Academic Press, London, pp. 301–320.

SUTHERLAND, AM, JACK, MA and LAVER, J (1988), 'Improved pitch detection algorithm employing temporal structure investigation of the speech waveform', *IEE Proc. Pt. F*, Vol. 135, No. 2, pp. 169–174.

SYRDAL, AK and GOPAL, HS (1986), 'A perceptual model of vowel recognition basd on the auditory representation of American English vowels', *J. Acoust. Soc. Am.*, Vol. 79, No. 4, pp. 1086–1100.

TAKAHASHI, K (1988), 'Transmission quality of evolving telephone services', *IEEE Trans. Commun.*, Vol. 26, No. 10, pp. 24–35.

THOMPSON, AR, MORAN, JM and SWENSON, GW (1986), *Interferometry and Synthesis in Radio Astronomy*, John Wiley, New York.

THOMSON, AH, PRATT, C and SIMPSON, H (1987), 'Nocturnal cough in asthma', *Arch. Dis. Child.*, Vol. 62, pp. 1001–1004.

THORNETT, C (1989), 'Technical aids for the disabled', *IEE Review*, Vol. 35, No. 5, pp. 165–168.

THORPE, CW and BATES, RHT (19XX), 'Speech analysis/companding/resynthesis by shift-and-add and Clean', *IEEE Trans. ASSP*, In revision.

THORPE, CW and DAWSON, SM (19XX), 'Automatic measurement of descriptive features of Hector's dolphin sonar signals', *J. Acoust. Soc. Am.*, To appear.

THORPE, CW, BRIESEMAN, NP, SQUIRES, PL and BATES, RHT (1986), 'Estimating glottal excitation by digital processing of recorded speech', In *New Zealand Medical Journal*, Proceedings of the Christchurch Medical Research Society, 16 July 1986, November, pp. 910–911.

THORPE, CW, FRIGHT, WR, TOOP, LJ and DAWSON, KP (19XXa), 'A micro-computer based interactive cough sound analysis system', *Computer Methods and Programs in Biomedicine*, Submitted for Publication.

THORPE, CW, BATES, RHT and DAWSON, SM (19XXb), 'Intrinsic echolocation capability of Hector's dolphin, *Cephalorhynchus hectori*', *J. Acoust. Soc. Am.*, In revision.

TITZE, I, BAER, T, COOPER, D and SCHERER, R (1983), 'Automated extraction of glottographic waveform parameters and regression to acoustic and physiologic variables', In BLESS, DM and ABBS, JH (Eds.), *Vocal Fold Physiology*, College-Hill Press, San Diego, Chap. 11, pp. 146–154.

TOBIAS, JV (Ed.) (1972), *Foundations of Modern Auditory Theory*, Vol. 2, Academic Press, New York.

TOHKURA, Y (1987), 'A weighted cepstral distance measure for speech recognition', *IEEE Trans. ASSP*, Vol. 35, No. 10, pp. 1414–1422.

TOOP, LJ (1989). Personal Communication.

TOOP, LJ, HOWIE, JGR and PAXTON, FM (1986), 'Night cough and general practice research', *J. R. Coll. Gen. Pract.*, Vol. 36, pp. 74–77.

TOOP, LJ, THORPE, CW and FRIGHT, WR (1989a), 'Cough sound analysis: A new tool for the diagnosis of asthma?', *Family Practice*, Vol. 6, No. 2, pp. 126–128.

TOOP, LJ, DAWSON, KP and THORPE, CW (1989b), 'Spectrographic analysis of cough sounds in asthma', *European Annals of Allergy and Clinical Immunology*, Vol. Supplement 5, p. 14.

TRAUNMÜLLER and BRANDERUD (1989), 'Paralinguistic speech signal transformations', *STL-QPSR*, No. 1/1989, pp. 63–68.

TRIBOLET, JM and CROCHIERE, RE (1979), 'Frequency domain coding of speech', *IEEE Trans. ASSP*, Vol. 27, No. 5, pp. 512–530.

TSANAKAS, JN, MILNER, RDG, MANNISTER, OM and BOON, AW (1988), 'The running asthma screening test', *Arch. Dis. Child.*, Vol. 63, pp. 261–265.

TUCKER, WR and BATES, RHT (1978), 'A pitch estimation alogrithm for speech and music', *IEEE Trans. ASSP*, Vol. 26, pp. 597–604.

TUCKER, WH, BATES, RHT, FRYKBERG, SD, HOWARTH, RJ, KENNEDY, WK, LAMB, MR and VAUGHAN, RG (1977), 'An interactive aid for musicians', *Int. J. of Man-Machine Studies*, Vol. 9, pp. 635–651.

TUKEY, PA (1983), 'Graphical methods', In GNANADESIKAN, R (Ed.), *Statistical Data Analysis*, American Mathematical Society, Providence, R.I., Chap. 2, pp. 8–48.

TURNER, SG (1986), *Real-time Speech Analysis for use with Impaired Speech Aids*, Master's thesis, Electrical and Electronic Engineering, University of Canterbury, NZ, March.

UN, CK and LEE, JR (1984), 'A 9600 bit/s RELP vocoder with split-band coding', In DEWILDE, P and MAY, CA (Eds.), *Int. Conf. on Commun.*, Elsevier, Holland, pp. 1174–1178.

VAISSIÈRE, J (1985), 'Speech recognition: a tutorial', In FALLSIDE, F and WOODS, WA (Eds.), *Computer Speech Processing*, Prentice Hall, New Jersey, Chap. 8.

VEENEMAN, DE and BEMENT, SL (1985), 'Automatic glottal inverse filtering from speech and electroglottographic signals', *IEEE Trans. ASSP*, Vol. 33, No. 2, pp. 369–377.

WAIBEL, A, HANAZAWA, T, HINTON, G, SHIKANO, K and LANG, KJ (1989), 'Phoneme recognition using time-delay neural networks', *IEEE Trans. ASSP*, Vol. 37, No. 3, pp. 328–339.

WAKITA, H (1973), 'Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms', *IEEE Trans. Audio Electroacoust.*, Vol. 21, No. 5, pp. 417–427.

WATKINS, WA (1974), 'Bandwidth limitations and analysis of cetacean sounds', *J. Acoust. Soc. Am.*, Vol. 55, No. 4, pp. 849–854.

WATKINS, WA (1979), 'Acoustics and the behavior of sperm whales', In BUSNEL, R and FISH, JF (Eds.), *Animal Sonar Systems*, Plenum Press, New York, pp. 283–289.

WATKINS, WA and WARTZOK, D (1985), 'Sensory biophysics of marine mammals', *Marine Mammal Science*, Vol. 1, No. 3, pp. 219–260.

WATKINS, WA, SCHEVILL, WE and BEST, PB (1977), 'Underwater sounds of *Cephalorhynchus heavisidii* (Mammalia: Cetacea)', *Journal of Mammalogy*, Vol. 58, No. 3, pp. 316–320.

WATSON, CI (1989). Personal Communication.

WATSON, CI, CLARK, TM, ELDER, AG and THORPE, CW (1988), 'Multifarious real-time speech processing applications', In *Proc. National Electronics Conference*, Christchurch, pp. 65–70.

WEBSTER, PM, SAWATZKY, RP, HOFFSTEIN, V, LEBLANC, R, HINCHEY, MJ and SULLIVAN, PA (1985), 'Wall motion in expiratory flow limitation: choke and flutter', *J. Appl. Physiol.*, Vol. 59, No. 4, pp. 1304–1312.

WENDLER, J, RAUHUT, A and KRÜGER, H (1986), 'Classification of voice qualities', *J. Phon.*, Vol. 14, pp. 483–488.

WICKELGREN, WA (1976), 'Phonetic coding and serial order', In CARTERETTE, EC and FRIEDMAN, MP (Eds.), *Handbook of Perception, Volume VII: Language and Speech*, Academic Press, New York, Chap. 7.

WIENER, N (1949), *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, John Wiley, New York.

WIERSMA, H (1982), 'Investigations on cetacean sonar IV, A comparison of wave shapes of Odontocete sonar signals', *Aquatic Mammals*, Vol. 9, No. 2, pp. 57–66.

WILKINSON, L (1987), *Systat: The System for Statistics*, Systat Inc., Evanston, IL.

WILLIAMS, CE and STAVENS, KN (1972), 'Emotions and speech: Some acoustical correlates', *J. Acoust. Soc. Am.*, Vol. 52, No. 4 (Pt. 2), pp. 1238–1250.

WIRTH, N (1983), *Programming in MODULA-2*, Springer-Verlag, Berlin, 2nd ed.

WITTEN, IH (1982), *Principles of Computer Speech*, Academic Press, London.

WODICKA, GR, STEVENS, KN, GOLUB, HL, CRAVALHO, EG and SHANNON, DC (1989), 'A model of acoustic transmission in the respiratory system', *IEEE Trans. Biomed. Eng.*, Vol. 36, No. 9, pp. 925–934.

WONG, DY, MARKEL, JD and GRAY, JR., AH (1979), 'Least squares glottal inverse filtering from the acoustic speech waveform', *IEEE Trans. ASSP*, Vol. 27, No. 4, pp. 350–355.

WONG, DY, JUANG, B and GRAY, JR., AH (1982), 'An 800 bit/s vector quantisation LPC vocoder', *IEEE Trans. ASSP*, Vol. 30, No. 5, pp. 770–780.

WOOD, LC and PEARCE, DJB (1989), 'Excitation synchronous formant analysis', *IEE Proc. Pt. I*, Vol. 136, No. 2, pp. 110–118.

WOOD, LC and TREITEL, S (1975), 'Seismic signal processing', *Proc. IEEE*, Vol. 63, No. 4, pp. 649–661.

WOODS, WA (1985), 'Language processing for speech understanding', In FALLSIDE, F and WOODS, WA (Eds.), *Computer Speech Processing*, Prentice Hall, New Jersey, Chap. 12, pp. 305–334.

WOODWARD, PM (1953), *Probability and Information Theory with Applications to Radar*, Pergamon.

YAO, K and THOMAS, JB (1967), 'On some stability and interpolatory properties of nonuniform sampling expansions', *IEEE Trans. Circuit Th.*, Vol. 14, No. 4, pp. 404–408.

YUEN, CK (1979), *Digital Spectral Analysis*, CSIRO, Australia. (D. Fraser, associate author).